2012

Statistical analysis of RNA-seq data from next-generation sequencing technology

Yaqing Si
Iowa State University

Statistical analysis of RNA-seq data from next-generation sequencing technology

by

Yaqing Si

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:

Peng Liu, Major Professor

Song X. Chen

Susan J. Lamont

Daniel S. Nettleton

Stephen B. Vardeman

Iowa State University

Ames, Iowa

2012

DEDICATION

To my parents

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, Dr. Peng Liu for her guidance, patience and support throughout this research and the writing of this thesis. Her insights and words of encouragement have often inspired me and renewed my hopes for completing my graduate education. I would also like to thank my committee members for their efforts and contributions to this work: Dr. Song X. Chen, Dr. Susan J. Lamont, Dr. Daniel S. Nettleton and Dr. Stephen B. Vardeman.

ABSTRACT

In recent years, the advent of next-generation sequencing (NGS) technology has been revolutionizing how genomic studies are processed. One important application of NGS technology is the study of transcriptome through sequencing of RNAs (RNA-seq). Compared with previous technologies such as microarray, RNA-seq data have many advantages, such as providing digital rather than analog signals of expression levels, dynamic and wider ranges of measurements, less noise, higher throughput, etc. Hence, RNA-seq is gradually replacing the array-based approach as the major platform in transcriptome studies. Meanwhile, the massive amounts of discrete data generated by the NGS technology call for effective methods of statistical analysis. There are many interesting questions in RNA-seq data analysis, and we focus on three important ones in this dissertation: identifying differentially expressed genes, from two-treatment experiments, detecting alternative splicing patterns using exon-expression data, and clustering gene expression profiles for multi-sample studies. Our major work are introduced in the following chapters:

First, we propose an approximated maximum-average powerful (AMAP) testing procedure to compare gene expression from two treatment groups. The proposed method allows for testing null hypotheses that are much more general than what have been considered by most previous studies, and it leads to a natural way of controlling the FDR. We show that our method has higher power as well as better FDR control than other widely-used methods in practice.

Second, we generalize the AMAP test from testing gene expression data to studying alternative splicing events from exon-level expression data. A nonparametric algorithm to estimate the distribution of exon usages is proposed, and this algorithm provides more flexibility for fitting the data, and higher computation efficiency. Our method is compared with previous methods and ours is shown to be much more powerful.

In the third project, we introduce clustering algorithms based on appropriate probability

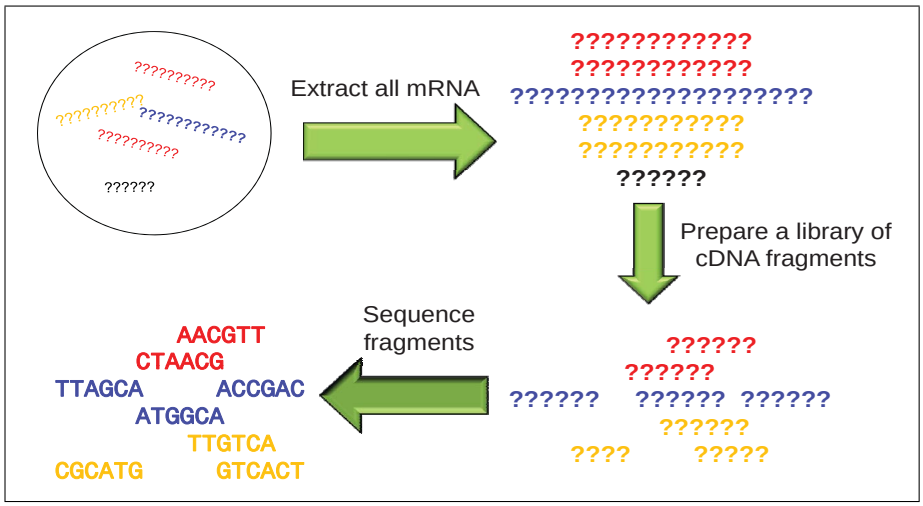
models for RNA-seq data, with well-designed initialization strategy and grouping algorithms. We also present a model-based hybrid-hierarchical clustering method to generate a tree structure that allows visualization of relationships among clusters as well as flexibility of choosing the number of clusters. Results from both simulation studies and analysis of a maize RNA-seq data set show that our proposed methods provide better clustering results than alternative methods that are not based on probability models.

CHAPTER 1. General Introduction

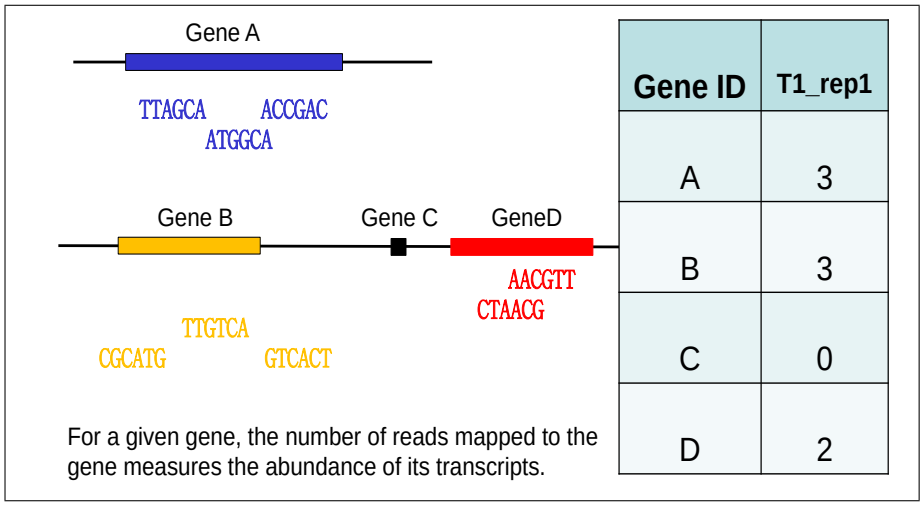
1.1 Next-Generation Sequencing Technology and RNA-seq Data

The recent advent of next-generation sequencing (NGS) technology has revolutionized genomic studies. One important application of NGS technology is to study transcriptome through sequencing of RNAs (RNA-seq). In a typical RNA-seq experiment, as shown in Figure 1.1, a sample of RNA is converted to a library of complementary DNA (cDNA) fragments and then sequenced on a high-throughput sequencing platform, such as Illumina Genome Analyzer, SOLiD or Roche 454 (Shendure and Ji, 2008). Millions of short sequences, or namely the *reads*, are obtained from this sequencing and then mapped to a reference genome or transcriptome, then the unmapped reads are usually discarded and mapped reads for each sample are assembled into gene-level, exon-level or transcript-level expression summaries, depending on the aims of the experiment, and the count of reads mapped to a given gene/exon/transcript measures the expression level for this region of the genome or transcriptome. See Table 1.1 for an example of a typical RNA-seq data set. In the remainder of thesis, we will use 'gene' as a general term for gene/transcript/exon, except in Chapter 3 which will be specified for studying exon-level RNA-seq data.

Compared with microarray, which has been the dominant approach of studying gene expression in the last two decades, RNA-seq technology has a wider measurable range of expression levels, less noise, higher throughput, and more information to detect allele-specific expression, novel promoters, and isoforms (Wang, Li and Brutnell, 2010; Oshlack et. al, 2010). For these reasons, RNA-seq is gradually replacing the array-based approach as the major platform in gene expression studies. Meanwhile, the massive amounts of discrete data generated by the NGS technology call for effective methods of statistical analysis. The challenging features of



(a) The sequencing step



(b) The mapping step

Figure 1.1: *The schematic procedures to obtain RNA-seq data.* (a) The sequencing step: a sample of mRNA is converted to a library of cDNA fragments and then sequenced on a high-throughput sequencing platform. Millions of short sequences, or namely the *reads*, are obtained. For simplicity, each read has length of 6 nucleotides on the figure, while in reality, the length varies from 26 to hundreds depending on sequencing platforms (Metzker, 2010). (b) The mapping step: the reads are mapped to a reference genome, and the mapped reads are counted for each gene to measures its expression level. Figures are adapted from lecture notes prepared by Dr. Peng Liu for Stat 416 class, ISU 2012.

Gene ID (Name) g	Gene Length L_g	Number of Mapped Reads					
		Treatment 1			Treatment 2		
		N_{g11}	N_{g12}	N_{g13}	N_{g21}	N_{g22}	N_{g23}
AC233926.1FG3	233	52	80	60	40	45	59
AC234179.1FG1	84	0	0	3	150	92	318
AF466202.2FG4	120	2	2	1	0	0	0
GRMZM2G0423	1304	177	382	200	10	7	6
GRMZM2G0056	587	1	12	7	20	12	38
...	...	...	...	...	...	...	...

Table 1.1: *A snapshot of a real RNA-seq data set.* The experiment has two treatments and each treatment has three replicates.

RNA-seq data include but mot limited to the following:

- *large number of genes*: there are often tens of thousands of genes to be compared simultaneously, hence researchers are more concerned about the overall performance of the analysis than that of a single gene. For example, we want to detect as many truly differentially expressed genes as possible while controlling multiple testing errors . The huge size of RNA-seq data set also requires intensive computation in analysis. Hence high computing power from both machine hardware and algorithm design are desired.

- *discrete data type*: RNA-seq data use counts of reads to quantify gene expressions, which are very different from continuous data that are can be conveniently modeled by Gaussian distributions. Though some data transformation technique, for instance, calculating the log-transformed counts, can be used to obtain continuous measurement of gene expressions, methods that keep and employ the nature of the count data are still preferred. In this sense, discrete probabilities, such as Poisson (Sultan et al., 2008), hypergeometric (Marioni et al., 2008), negative-binomial (NB) (Robinson and Oshlack, 2010; Anders and Huber, 2010). distributions have proposed to model the counts. However, difficulties often exist in the computation or knowing the properties of statistics based on these distributions.

1.2 Detecting Differentially Expressed Genes

One of the primary objectives for most RNA-seq experiments is to compare the gene expression levels across various treatments. A simple and common RNA-seq study involves two treatments in a randomized complete design , for example, treated versus untreated cells, two different tissues from a mouse, cancer or heathy human beings, etc. In these studies, researchers are particularly interested in detecting gene with differential expressions (DE), i.e., genes whose expression levels differ between the two treatments. Detecting DE genes can also be an important pre-step for subsequent studies, such as clustering gene expression profiles or testing gene set enrichments.

Several methods have been proposed for detecting DE genes based on RNA-seq data. Among them, Fisher's exact test (Bloom et al., 2009), χ^2 goodness-of-fit test (Marioni et al., 2008), likelihood ratio test (LRT) (Bullard et al., 2010) and the PoissonSeq procedure (Li et al., 2011) are based on Poisson models for the count data, mostly from RNA-seq experiments that use only technical replicates. However, when there are biological replicates, RNA-seq data may exhibit more variability than what the Poisson distribution predicts, and then the negative binomial (NB) distribution has been used to model the counts in such cases. Based on NB models, several tests have been developed and implemented in the R packages, for instance, `edgeR` (Robinson and Smyth, 2008), `DESeq` (Anders and Huber, 2010) and `baySeq` (Hardcastle and Kelly, 2010), etc.

Although the above-mentioned methods have been proposed to detect DE genes, there are no theoretical justifications for whether any of these methods are optimal or how to search for the optimal test. Furthermore, most proposed tests are designed for testing whether the mean expression levels are exactly the same or not across treatments, whereas, sometimes, biologists are interested in detecting genes with expression changes larger than a certain threshold. Another issue with current methods is that the multiple testing errors are not well studied. Currently widely-used procedures, include those proposed by Benjamini and Hochberg (1995) and Storey and Tibshirani (2003) for testing methods that generate p-values, are often found to control the false discovery rate (FDR) either conservatively or literally (Li et al., 2011; Kvam

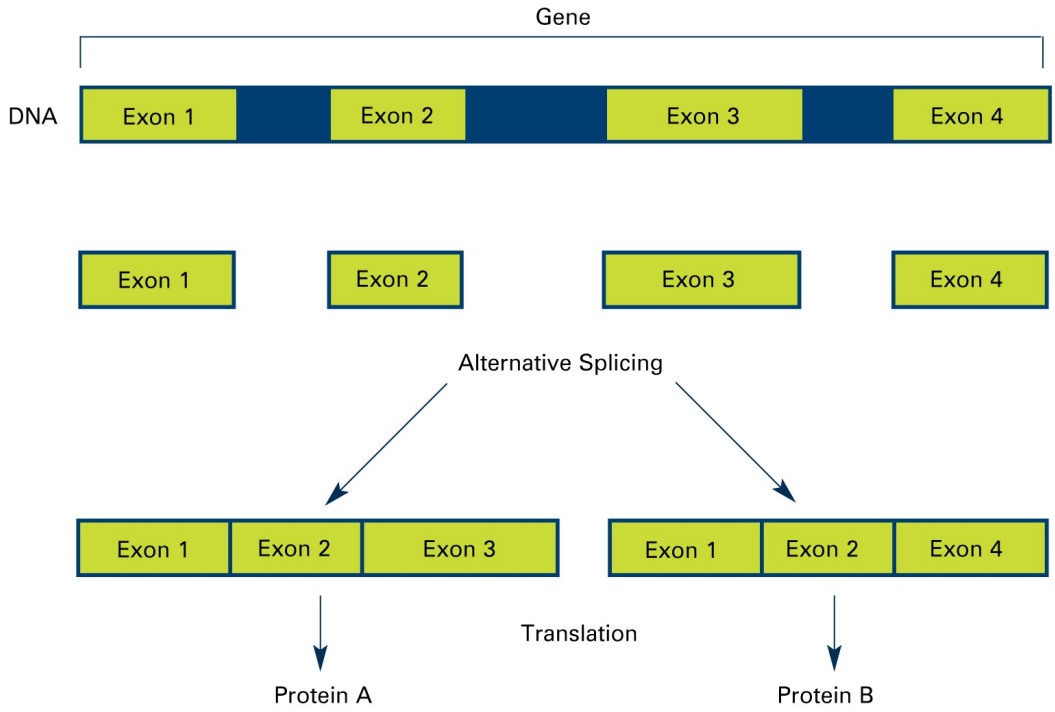


Figure 1.2: *Alternative splicing*. Two isoforms from one gene: exon 4 is skipped to produce protein A, and exon 3 is skipped for protein B. Figure is from http://images.nigms.nih.gov.

et al., 2012). Hence a new and better performing method of testing for DE genes is in high demand.

We propose an approximated maximum-average powerful (AMAP) testing methods to compare gene expressions from two treatment groups. The proposed method allows for testing null hypotheses that are much more general than what have been considered by most previous studies, and it leads to a natural way of controlling the FDR. We show that our method has higher power as well as better FDR control than other widely-used methods in practice.

1.3 Alternative Splicing

For eukaryotic cells, it is common that a gene has several protein-coding regions called *exons*, and the exons of a gene are reconnected in multiple ways during RNA splicing. The resulting different mRNAs are translated into different protein isoforms (see Figure 1.2 for the illustration). This process is called *alternative splicing* (AS). AS affects message stability and translation efficiency, and increases protein diversity (Black, 2003; Stamm et al., 2005). AS

in particular is known to affect more than half of all human genes, and has been proposed as a primary driver of the evolution of phenotypic complexity in mammals (Lander et al., 2001; Johnson et al., 2003). So studying AS events has been an important question for scientists.

With RNA-seq data, AS can be studied by comparing the coverages of the exons from different treatments. Though some techniques as introduced in section 1.2 to detect differential gene expressions can also be used in comparing exon coverages, specific tools to analyze exon coverages are still very limited to our best knowledge. Some available methods such like `DEXSeq` (Anders et al., 2012) and `MATS` (Shen et al., 2012) have been developed to test for differential exon usages. However, the detection power of these tests has not been evaluated very well. Moreover, some other interesting AS patterns still need more investigation. For example, biologists are often interested in testing for expressed exons, or the extreme 'switch-like' pattern, which means that the exon is expressed in one treatment but not in another (see Figure 1.2 where exon 3 is only expressed in the first treatment to produce protein A). All these indicate that studying AS is challenging as well as full of opportunities for statisticians.

We generalize the AMAP test from testing gene expression data to studying alternative splicing events from exon-level expressions. A nonparametric algorithm to estimate the distribution of exon usages is proposed, and this algorithm provides more flexibility for fitting the data, and higher efficiency of computation. Our methods is compared with previous methods and is shown to be much more powerful.

1.4 Cluster Analysis

Some RNA-seq experiments involve more than two treatment groups. For example, Li et al. (2010) measured the gene expressions from four representative sections of a leaf blade from a corn plant. By surveying the gene expression profiles along different developmental stages of the leaf, the transcriptional network associated with the development of C4 photosynthesis can be understood. Similar to in microarray studies, in these sequencing experiments with multiple treatment groups, cluster analysis groups genes with similar expression patterns across the treatments. And because genes within such groups often tend to be functionally related, cluster analysis has been employed as an important technique to provide insight into gene functions

and networks.

Many heuristic algorithms, such as K-means and self-organizing map (SOM) (Xiao et al., 2003), have been popularly applied to microarray analysis, and they can potentially be applied to RNA-seq data indirectly, for example, to log-transformed or Z-scores of RPKM of the read counts. However, studies of clustering algorithms with microarray data already revealed that heuristic algorithms usually perform worse than model-based algorithms (Yeung et al, 2001). Hence it is desirable to develop a clustering algorithm based on appropriate probability models specially for RNA-seq data and enhance the performance.

We introduce clustering algorithms based on appropriate probability models for RNA-seq data, with well-designed initialization strategy and grouping algorithms. We also present a model-based hybrid-hierarchical clustering method to generate a tree structure that allows visualization of relationships among clusters as well as flexibility of choosing the number of clusters. Results from both simulation studies and analysis of a maize RNA-seq data set show that our proposed methods provide better clustering results than alternative methods that are not based on probability models.

1.5 Dissertation Organization

The main chapters of this dissertation focus on the three questions introduced in section 1.2, 1.3 and 1.4: In Chapter 2, we present the AMAP test to compare gene expression levels from two treatment groups; In Chapter 3, the AMAP test is generalized to studying alternative splicing events from exon-level expression data; the model-based clustering method for RNA-seq data is introduced in Chapter 4. A summary about our work and possible directions for future research are briefly discussed in Chapter 5.

CHAPTER 2. An Optimal Test with Maximum Average Power While Controlling FDR with Application to RNA-seq Data

Yaqing Si and Peng Liu

Iowa State University, Snedecor Hall, Ames, IA 50011, USA.

This work was submitted to *Biometrics*

Abstract

The recent RNA-seq technology is an attractive method to study gene expression. One of the most important goals in RNA-seq data analysis is to detect genes differentially expressed across treatments. Although several statistical methods have been published, there are no theoretical justifications for whether these methods are optimal or how to search for the optimal test. Furthermore, most proposed tests are designed for testing whether the mean expression levels are exactly the same or not across treatments, whereas, sometimes, biologists are interested in detecting genes with expression changes larger than a certain threshold. Another issue with current methods is that the false discovery rate (FDR) control is not well studied. In this manuscript, we proposed a test to address all above issues. Under model assumptions, we derive an optimal test that achieves the maximum of average power among those that control FDR at the same level. We also provide an approximated version, the approximated most average powerful (AMAP) *test, for practical implementation. The proposed method allows for testing null hypotheses that are much more general than what have been considered by most previous studies, and it leads to a natural way of controlling the FDR. Through simulation studies, we show that our test has higher power than other methods, including the widely-used edgeR, DESeq, and baySeq methods, as well as better FDR control than two other FDR control*

procedures commonly used in practice. For demonstration, we also apply the proposed method to a real RNA-seq dataset obtained from maize.

Key Words: Empirical Bayes; FDR control; Gene expression; Maximum average power; RNA-seq.

2.1 Introduction

The recent advent of next-generation sequencing (NGS) technology has revolutionized genomic studies. One important application of NGS technology is the study of the transcriptome through sequencing of RNAs (RNA-seq). In a typical RNA-seq experiment, a sample of RNA is converted to a library of complementary DNA fragments and then sequenced on a high-throughput sequencing platform, such as Illumina's Genome Analyzer. Millions of short sequences, or *reads*, are obtained from this sequencing and then mapped to a reference genome. The count of reads mapped to a given gene measures the expression level of this gene. In the last two decades, microarray technology has been the dominant approach of studying gene expression. Compared with microarray, RNA-seq technology has a wider measurable range of expression levels, less noise, higher throughput, and more information to detect allele-specific expression, novel promoters, and isoforms (Wang, Li and Brutnell, 2010; Oshlack et. al, 2010). For these reasons, RNA-seq is gradually replacing the array-based approach as the major platform in gene expression studies. Meanwhile, the massive amounts of discrete data generated by the NGS technology call for effective methods of statistical analysis. For statistical analysis of RNA-seq data, detecting differentially expressed (DE) genes across treatments/conditions is essential, and commonly the major goal of the analysis of RNA-seq data. Additionally, detecting DE genes can be a pre-step for subsequent studies, such as clustering gene expression profiles or testing gene set enrichments. In this paper, we focus on the question of detecting DE genes from RNA-seq data. The proposed method can also be applied to the analysis of other types of NGS data, for instance, chromatin immunoprecipitation-sequencing (ChIP-seq) data.

Several methods have been proposed for detecting DE genes based on RNA-seq data. Two popular distributions for fitting RNA-seq data are the Poisson and negative binomial (NB)

distributions. Early RNA-seq studies that used only technical replicates reported that Poisson distributions fit well to the counts for the majority of genes (Marioni et al., 2008; Bullard et al., 2010). Fisher's exact test (Bloom et al., 2009), χ^2 goodness-of-fit test (Marioni et al., 2008), likelihood ratio test (LRT) (Bullard et al., 2010), and the PoissonSeq procedure (Li et al., 2011) were applied to detect DE genes. However, when there are biological replicates, RNA-seq data may exhibit more variability than what the Poisson distribution predicts, i.e., the variances are likely greater than the means for a considerable number of genes (Anders and Huber, 2010). This phenomenon is called over-dispersion. In such cases, the NB distribution, which allows the variance to exceed the mean, has been used to model the counts. Based on NB models, several tests have been developed and implemented in the R packages edgeR (Robinson and Smyth, 2008), DESeq (Anders and Huber, 2010) and baySeq (Hardcastle and Kelly, 2010).

Because RNA-seq experiments are still expensive, such experiments typically involve only a few samples from each treatment group. However, each experiment measures an enormous number of genes. For example, the number of measured genes is more than 30,000 for human beings (Pickrell et. al, 2010), and more than 50,000 for maize (Li et al., 2010). This results in the "large p, small n" problem for detecting DE genes. A few methods have been proposed to borrow information across genes in order to achieve better performance in the multiple testing procedure. For example, Robinson and Smyth (2007) proposed an estimator for the dispersion parameter in the NB model that shrinks the dispersion parameter for each individual gene toward a common value using the weighted likelihood approach; Anders and Huber (2010) proposed a local regression model of the variance on the mean using all genes, giving a fitted relationship useful for estimating dispersion parameters; Hardcastle and Kelly (2010) proposed the baySeq method, a test with an empirical Bayes approach. All these methods show higher detection power than those that do not share information. However, there is no theoretical justification for the optimality of these existing methods and no discussion on how to search for the optimal test for RNA-seq data.

In addition to the lack of theoretical guidance in deriving the optimal test, the false discovery rate (FDR) control is not well studied. FDR has been widely applied to multiple testing problems encountered in RNA-seq experiments and other genomic studies. With currently

available methods, one may control the FDR with the Benjamini and Hochberg's procedure (Benjamini and Hochberg, 1995) or Storey and Tibshirani's procedure (Storey and Tibshirani, 2003) after obtaining the p-values from a test. Only a few studies investigated the performance of these methods in the context of RNA-seq data analysis. Simulation studies by Li et al. (2011) suggest that the FDR control of edgeR may be conservative sometimes. Kvam et al. (2012) also showed that applying Benjamini and Hochberg's procedure to the p-values generated by DESeq or edgeR was conservative in some cases while liberal in some other cases. Our simulation results support the same conclusion (see Figure 2.7). Because FDR control directly affects the final list of DE genes declared by the tests, a good FDR control procedure is highly desired.

Moreover, most existing tests are designed for testing the null hypothesis that the difference between expression levels of different treatments is exactly zero for each gene. However, a slight change in the mean expression levels may not be biologically significant. Sometimes, it is more valuable to detect genes with big changes in the mean expressions (MacCarthy and Smyth, 2009; Peart et. al, 2005; Covshoff et al., 2008). A common strategy is to select a list of genes with both high statistical significance and large fold-changes (FC) of expressions between treatments without further evaluation of the FDR of the resulting list (Peart et. al, 2005; Covshoff et al., 2008).

In this paper, we develop an optimal test for RNA-seq data analysis while controlling the FDR, where the optimality is defined as achieving the maximum of the power averaged across all genes for which null hypotheses are false. We call such test the maximum average power (MAP) test, a concept introduced in Chen et al. (2007) and further studied in Hwang and Liu (2010) for microarray data analysis. The statistic of the proposed test provides a natural way of controlling the FDR. Furthermore, the null hypothesis flexibly adapts to the context of problem. The MAP tests are derived for both Poisson and NB distributed data, respectively, where the parameters of these distributions are assumed to come from appropriate hyper distributions. In practice, we do not know the "true" hyper distributions, and we propose to estimate them with mixture distributions. Plugging in the estimated hyper distribution leads to an approximated MAP (AMAP) test. We perform a variety of simulation studies using the Poisson distribution, NB distribution, and real RNA-seq data. Simulation results show that the AMAP test performs

numerically indistinguishable to the optimal MAP test for Poisson data, and the performance of the AMAP test is close to that of the MAP test for NB data. The AMAP test outperforms Fisher's exact test, edgeR, DESeq and baySeq in most simulation settings. In addition, the results demonstrate that our method provides accurate estimation of the FDR for both the AMAP test and the other tests.

This article is organized as follows: In section 2.2, we describe the proposed method for the Poisson model and for the NB model; In section 2.3, we simulate RNA-seq data using Poisson models, NB models and real data, respectively, under a variety of settings, and we evaluate the effectiveness of the proposed method and other methods; In section 2.4, we analyze a real dataset using our proposed methods and some existing methods; Section 2.5 provides some discussion.

2.2 Method

2.2.1 Poisson Model

Suppose that an RNA-seq dataset has G genes. Let X_{gij} denote the number of reads mapped to gene g from replicate j of treatment i, where $g = 1, \cdots, G$, $i = 1, 2$, $j = 1, \cdots, n_i$, and $n_i \geq 1$ is the number of replicates in treatment group i. Poisson distributions have previously been used to model the counts when there are only technical replicates of one biological sample for each treatment group (Bullard et al., 2010; Marioni et al., 2008). Assuming $X_{gij} \sim \text{Poisson}(\lambda_{gij})$, we model the mean of the Poisson distribution, λ_{gij}, as

$$\lambda_{gij} = S_{ij}\lambda_g \exp(\rho_i \delta_g), \tag{2.1}$$

where λ_g represents the overall geometric mean expression level of gene g across both treatments; $\rho_1 = -1/2$ and $\rho_2 = 1/2$ so that δ_g is the log fold change (log-FC) between the two treatment means; and S_{ij} is a normalization factor that adjusts for varying sequencing depths and potentially other technical effects across the replicates. Several proposed methods normalize RNA-seq data by estimating the normalization factor in different ways. For example, we can estimate S_{ij} by the total number of mappable reads (Mortazavi et al., 2008), the 75th percentile of the non-zero counts (Bullard et al., 2010), the median estimated from the count

ratio to a pseudoreference (Anders and Huber, 2010), or a scalar estimated by method of the trimmed mean of M-values (TMM) (Robinson and Oshlack, 2010). In Appendix 2.A.6, we revisit this issue and present results with different normalization methods. After estimation, S_{ij} is usually treated as known in the followup analysis.

2.2.2 Hypotheses

One major goal of RNA-seq experiments is to identify genes whose expression levels change across different treatment groups. To achieve this goal, we test the following hypotheses regarding the parameter δ_g for each gene g:

$$H_0^g : \delta_g \in \Delta_0 \text{ v.s. } H_1^g : \delta_g \in \Delta_1, \tag{2.2}$$

where Δ_0 and Δ_1 correspond to the null and alternative sets of values for δ_g, respectively, and they represent a partition of the real line $\mathcal{R}$. The null space Δ_0 can be defined in different ways depending on the biological questions of interest. For example, we set $\Delta_0 = \{0\}$ if we are interested in knowing whether the mean expression levels in the two treatments are equal. If we are interested in whether the mean expression is higher in the second treatment than in the first, we set $\Delta_0 = (-\infty, 0]$. Sometimes, biologists want to detect genes whose expression changes are large enough, for instance, with fold-changes of expressions greater than 1.5 (Peart et. al, 2005). In this case, we set $\Delta_0 = \{\delta : |\delta| \leq c\}$ with $c = \log 1.5$. The test we derive in the next section allows for any Δ_0 that is a subset of $\mathcal{R}$. Most other tests only allow the simple null hypothesis $\Delta_0 = \{0\}$. This is an important advantage of our method over others.

2.2.3 Test for the Poisson Model

Our goal is to derive MAP tests while controlling FDR. As in Storey (2007) and in Hwang and Liu (2010), we focus on the hypothesis rejecting strategy that does not depend on individual genes. Storey (2007) calls such type of testing procedure the single thresholding procedure (STP) and explains that it is often the only available option in practice. Theorem 3 in Hwang and Liu (2010) proves that, for STPs, the MAP test while controlling the average type I error

rate is the MAP test when controlling the FDR. Therefore, we will derive the MAP test by maximizing the average power while controlling the average type I error rate.

Let $\boldsymbol{X}_g = \{X_{gij} : i = 1, 2; j = 1, \cdots, n_i\} \in \mathcal{X}$ denote the vector of observations for gene g, where $\mathcal{X}$ is the data space that contains all possible values for $\boldsymbol{X}_g$. Let $f(\boldsymbol{X}_g|\lambda_g, \delta_g)$ be the likelihood function for the Poisson model (2.1), and let $\varphi(\boldsymbol{X}_g)$ be a critical function that takes the value of 1 or 0 so that the hypothesis H_0^g is rejected if and only if $\varphi(\boldsymbol{X}_g) = 1$. Then, the power for testing gene g is $\int_{\mathcal{X}} \varphi(\boldsymbol{X}_g) f(\boldsymbol{X}_g|\lambda_g, \delta_g) \mathrm{d}\boldsymbol{X}_g$, and the average power is $\frac{1}{G_1} \sum_{\{g:\delta_g \in \Delta_1\}} \int_{\mathcal{X}} \varphi(\boldsymbol{X}_g) f(\boldsymbol{X}_g|\lambda_g, \delta_g) \mathrm{d}\boldsymbol{X}_g$, where G_1 is the number of genes with $\delta_g \in \Delta_1$. When the (λ_g, δ_g)'s for $\delta_g \in \Delta_1$ are assumed to be random variables coming from a distribution with probability density function (PDF) $\pi_1(\lambda, \delta)$ with support $\mathcal{R}^+ \times \Delta_1$, then as $G_1 \to \infty$, the average power converges to $\int_{\mathcal{R}^+} \int_{\Delta_1} \left[\int_{\mathcal{X}} \varphi(\boldsymbol{X}) f(\boldsymbol{X}|\lambda, \delta) \mathrm{d}\boldsymbol{X} \right] \pi_1(\lambda, \delta) \mathrm{d}\delta \mathrm{d}\lambda$, which by Fubini's theorem is equal to

$$\int_{\mathcal{X}} \varphi(\boldsymbol{X}) \left[\int_{\mathcal{R}^+} \int_{\Delta_1} f(\boldsymbol{X}|\lambda, \delta) \pi_1(\lambda, \delta) \mathrm{d}\delta \mathrm{d}\lambda \right] \mathrm{d}\boldsymbol{X}. \tag{2.3}$$

Note that the subscript g is not needed in this integration. Similarly, if we assume the (λ_g, δ_g)'s of the genes with $\delta_g \in \Delta_0$ follow a distribution with PDF $\pi_0(\lambda, \delta)$ with support $\mathcal{R}^+ \times \Delta_0$, then the average type I error rate of these tests approaches

$$\int_{\mathcal{X}} \varphi(\boldsymbol{X}) \left[\int_{\mathcal{R}^+} \int_{\Delta_0} f(\boldsymbol{X}|\lambda, \delta) \pi_0(\lambda, \delta) \mathrm{d}\delta \mathrm{d}\lambda \right] \mathrm{d}\boldsymbol{X} \tag{2.4}$$

as $G_0 \to \infty$, where $G_0 = G - G_1$.

Applying the Neyman-Pearson lemma, we claim that the optimal test that maximizes the average power (2.3) while controlling the average type I error rate (2.4) rejects H_0^g when the following statistic is small:

$$T^*(\boldsymbol{X}_g) = \frac{\int_{\mathcal{R}^+} \int_{\Delta_0} f(\boldsymbol{X}_g|\lambda, \delta) \pi_0(\lambda, \delta) \mathrm{d}\delta \mathrm{d}\lambda}{\int_{\mathcal{R}^+} \int_{\Delta_1} f(\boldsymbol{X}_g|\lambda, \delta) \pi_1(\lambda, \delta) \mathrm{d}\delta \mathrm{d}\lambda}.$$

See the Appendix for the proof for this claim. If we define a mixture distribution of (λ_g, δ_g) by $\pi(\lambda, \delta) = p_0 \pi_0(\lambda, \delta) + (1 - p_0) \pi_1(\lambda, \delta)$, where p_0 is the proportion of genes with $\delta_g \in \Delta_0$, and apply a monotonic transformation of $T^*(\boldsymbol{X}_g)$, we obtain an equivalent test statistic:

$$T(\boldsymbol{X}_g) = \frac{\int_{\mathcal{R}^+} \int_{\Delta_0} f(\boldsymbol{X}_g|\lambda, \delta) \pi(\lambda, \delta) \mathrm{d}\delta \mathrm{d}\lambda}{\int_{\mathcal{R}^+} \int_{\mathcal{R}} f(\boldsymbol{X}_g|\lambda, \delta) \pi(\lambda, \delta) \mathrm{d}\delta \mathrm{d}\lambda}. \tag{2.5}$$

Applying Theorem 3 of Hwang and Liu (2010), we can also prove that the test that rejects H_0^g when the statistic (2.5) is small also maximizes the average power among the tests that control FDR at the same level. Now we formally summarize this result in the following theorem:

Theorem 1. Most Average Powerful (MAP) Test. *The test that maximizes the average power (2.3) with FDR controlled at level α is the test that rejects H_0^g using the rejection region*

$$\mathcal{C} = \{\boldsymbol{X}_g : T(\boldsymbol{X}_g) \leq c\}$$

for $g = 1, 2, ..., G$, where the test statistic $T(\boldsymbol{X}_g)$ is defined in (2.5), and the constant c is the critical value so that the multiple testing procedure has FDR controlled at level α.

We call the test with critical function $\varphi^*(\boldsymbol{X}) := I(\boldsymbol{X} \in \mathcal{C})$ as described in Theorem 1 an *MAP* test, where $I(\cdot)$ is the indicator function.

2.2.4 FDR Control

In this section, we present how we estimate the FDR level of the MAP test given a critical value for the test statistic. One can control the FDR to the desired level by choosing the appropriate critical value.

Within Bayesian framework, $\pi(\lambda, \delta)$ can be considered as the prior distribution for (λ_g, δ_g). It is straightforward to show that $T(\boldsymbol{X}_g)$ defined in (2.5) is $\mathrm{P}(\delta_g \in \Delta_0 | \boldsymbol{X}_g)$, the posterior probability of $\delta_g \in \Delta_0$ given the data. Then, for a test with critical function $\varphi(\boldsymbol{X}_g)$, the expected number of false positives (EFP) can be estimated by $\sum_g T(\boldsymbol{X}_g)\varphi(\boldsymbol{X}_g)$. Finally the FDR can be estimated by the ratio of the estimated EFP to the number of rejected hypotheses:

$$\widehat{\mathrm{FDR}} = \frac{\sum_g T(\boldsymbol{X}_g)\varphi(\boldsymbol{X}_g)}{\sum_g \varphi(\boldsymbol{X}_g)}, \tag{2.6}$$

which can be viewed as the *average posterior probability of being false positive* for the list of genes declared to be positives by φ. Note that the above critical function $\varphi(\boldsymbol{X}_g)$ is *not* necessarily for the MAP test φ^*, but can be for *any* test that is an STP. As a result, the FDR levels for all tests introduced in section 2.1 can be estimated by equation (2.6).

2.2.5 Approximation of $\pi(\lambda, \delta)$ and the Resulting AMAP Test

The derivation of the MAP test assumes the knowledge of the joint distribution $\pi(\lambda, \delta)$. However, in practice, we do not know this distribution and need to estimate it. Considering the high dimensionality of tests, computational efficiency is highly desired. For this reason, we assumes a parametric model for the distribution of $\pi(\lambda, \delta)$. In addition, we would like our model to be flexible enough to provide good fitting for various datasets. The model we propose for $\pi(\lambda, \delta)$ is:

$$\sum_{k=1}^{K} q_k G(\lambda|\alpha_k, \beta_k) N(\delta|\mu_k, \sigma_k), \tag{2.7}$$

where K is the number of components of the mixture model; q_k is the weight of mixing component k with $q_k > 0$ and $\sum_{k=1}^{K} q_k = 1$, $G(\cdot\,|\alpha_k, \beta_k)$ is the PDF of a Gamma distribution that has mean α_k/β_k and variance α_k/β_k^2, and $N(\cdot\,|\mu_k, \sigma_k)$ is the PDF of a Normal distribution with mean μ_k and variance σ_k^2. We call the distribution (2.7) a K-component *mixture Gamma-Normal* (MGN) distribution. Note that the Gamma distribution is the conjugate prior for the Poisson distribution, which will simplify the calculation of $T(\boldsymbol{X}_g)$ in equation (2.5) by reducing one dimension of integration. In addition, by varying the number of components, the mixing weights, and the parameters of each component, the mixture distribution provides ample model flexibility.

Given a positive integer K, the unknown hyperparameters parameters in $\pi(\lambda, \delta)$ are

$$\boldsymbol{\theta} = \{(q_k, \alpha_k, \beta_k, \mu_k, \sigma_k) : k = 1, 2, \cdots, K\}.$$

In Appendix 2.A.2, we provide an Expectation-Maximization (EM) algorithm to estimate these parameters simultaneously. Plugging the estimated $\pi(\lambda, \delta)$ into the formula for the MAP statistic $T(\boldsymbol{X}_g)$, we obtain an approximated MAP (AMAP) test. To apply AMAP test in practice, we also need to determine a proper value for K.

If $K = 1$, λ_g and δ_g are assumed independent by model (2.7). This is not necessarily appropriate for some RNA-seq data. In addition, the one-component MGN distribution may not be able to provide a good approximation to the true distribution. Considering the mean expression λ only, the single Gamma distribution often expects fewer highly expressed genes than what is observed (Ji et al., 2008). With more components, model (2.7) allows dependence

between λ and δ, a more flexible shape for the MGN distribution, and hence a likely better approximation of the true distribution. However, a bigger K means more hyperparameters in $\pi(\lambda, \delta)$ to estimate as well as more computation to calculate the statistic $T(\boldsymbol{X}_g)$. Thus in practice, it is not desirable to choose a very large K.

Our experience with several datasets suggests that $K = 3$ usually provides a good performance. To illustrate how to choose K, we show an example of analyzing a real RNA-seq dataset from Sultan et al. (2008), which studied the transcriptomes from a human embryonic kidney and a B cell line without biological replicates. See Appendix 2.A.3 for the details of the analytical results. For this dataset, the model parameters of the MGN distribution for $K = 1, 2, \cdots, 10$ were estimated. From visual inspection of model-fit (Figure 2.5(a)-2.5(c)), comparison of the values of the AMAP statistics (Figure 2.5(d)-2.5(e)), and the Bayesian Information Criterion (BIC) scores calculated for these MGN-Poisson hierarchical models (Figure 2.5(f)), we found that $K = 3$ provides the best fit for this dataset.

The model (2.7) without parameter constraints works well when the null space Δ_0 consists of one or several intervals of $\mathcal{R}$. However, if we test for the simple null hypothesis $\delta_g = \delta_0$, it is found that $T(\boldsymbol{X}_g) \equiv 0$ according to equation (2.5) because of the continuity of Normal distributions. To solve this problem, we view the δ for the null genes as coming from a degenerated Normal distribution, $N(\delta_0, 0)$, that has a point mass at δ_0. Then, the joint distribution of λ and δ for the null genes takes the form of $\pi_0(\lambda, \delta_0) = \pi_0(\lambda)N(\delta_0, 0)$. To estimate $\pi(\lambda, \delta) = p_0\pi_0(\lambda)N(\delta_0, 0) + (1 - p_0)\pi_1(\lambda, \delta)$, we estimate $\pi_0(\lambda)$ by a K_0-component mixture Gamma distribution and $\pi_1(\lambda, \delta)$ by a K_1-component MGN distribution, where K_0 and K_1 are positive integers. Hence, we approximate $\pi(\lambda, \delta)$ by a $(K = K_0 + K_1)$-component MGN distribution with parameters $\{(q_k, \alpha_k, \beta_k, \mu_k, \sigma_k) : k = 1, \cdots, K\}$, among which K_0 components have known parameters $\mu_k = \delta_0$ and $\sigma_k = 0$ for $k = 1, \cdots, K_0$. All the unknown parameters can be estimated by the EM algorithm described in Appendix 2.A.2. We found that $K_0 = K_1 = 3$ works well for several real datasets we examined.

2.2.6 AMAP Test for the Negative-Binomial Model

As mentioned in section 2.1, RNA-seq data with biological replicates often exhibit over-dispersion while fitting the Poisson model. Assuming an NB instead of a Poisson model is one way to deal with over-dispersed data because the NB distribution specifies that the variance is greater than the mean. Following Robinson and Smyth (2007), we parameterize the variance of the NB distribution by:

$$\mathrm{Var}(X_{gij}) = \lambda_{gij} + \phi_g \lambda_{gij}^2, \tag{2.8}$$

where λ_{gij} is the mean and is modeling using (2.1), and ϕ_g is the dispersion parameter that determines the extra variability compared to the Poisson model. The variance $\mathrm{Var}(X_{gij})$ approaches the mean λ_{gij} when the dispersion parameter ϕ_g diminishes to 0, thus the Poisson model can be viewed as a special NB model that has zero dispersion (Robinson and Smyth, 2007).

The NB model defined by equations (2.1 & 2.8) has three unknown parameters, $(\lambda_g, \delta_g, \phi_g)$, for each gene. Assuming we know their joint distribution, $\pi(\lambda, \delta, \phi)$, the MAP test statistic takes the following form:

$$T(\boldsymbol{X}_g) = \frac{\int_{\mathcal{R}^+}\int_{\mathcal{R}^+}\int_{\Delta_0} f(\boldsymbol{X}_g|\lambda,\delta,\phi)\pi(\lambda,\delta,\phi)\mathrm{d}\delta\mathrm{d}\lambda\mathrm{d}\phi}{\int_{\mathcal{R}^+}\int_{\mathcal{R}^+}\int_{\mathcal{R}} f(\boldsymbol{X}_g|\lambda,\delta,\phi)\pi(\lambda,\delta,\phi)\mathrm{d}\delta\mathrm{d}\lambda\mathrm{d}\phi}. \tag{2.9}$$

Again, $\pi(\lambda, \delta, \phi)$ is unknown in practice. We could generalize the MGN model (2.7) for $\pi(\lambda, \delta)$ so that $\pi(\lambda, \delta, \phi)$ is approximated by the mixture model $\sum_{k=1}^{K} q_k G(\lambda|\alpha_k, \beta_k) N(\delta|\mu_k, \sigma_k) G(\phi|a_k, b_k)$ with additional hyperparameters $a_k, b_k > 0$. However, since there is no obvious conjugate prior for the NB model, the three-dimensional integrations in the calculation of the test statistic (2.9) and the EM algorithm for estimating the hyperparameters require intensive computation.

Instead of trying to estimate $\pi(\lambda, \delta, \phi)$ and compute the three-dimensional integrations, we take an approximating approach for the NB model. First, we estimate the dispersion parameter ϕ_g for each gene by the quasi-likelihood (QL) approach. Other methods of estimating ϕ_g such as those discussed in Nelder (2000) and Robinson and Smyth (2008) can also be applied. Then, the estimate $\widehat{\phi}_g$ is treated as the true ϕ_g for gene g. We model $\pi(\lambda, \delta)$ by an MGN distribution as described in section 2.2.5 and estimate the model parameters by the EM algorithm as described

in Appendix 2.A.2. With the estimated distribution, $\widehat{\pi}(\lambda, \delta)$, the AMAP statistic is

$$T(\boldsymbol{X}_g) = \frac{\int_{\mathcal{R}^+} \int_{\Delta_0} f(\boldsymbol{X}_g|\lambda, \delta, \widehat{\phi}_g)\widehat{\pi}(\lambda, \delta)\mathrm{d}\delta\mathrm{d}\lambda}{\int_{\mathcal{R}^+} \int_{\mathcal{R}} f(\boldsymbol{X}_g|\lambda, \delta, \widehat{\phi}_g)\widehat{\pi}(\lambda, \delta)\mathrm{d}\delta\mathrm{d}\lambda}, \tag{2.10}$$

where the likelihood function $f(\boldsymbol{X}_g|\lambda, \delta, \widehat{\phi}_g)$ is calculated based on the NB model (2.1 & 2.8). The FDR for the AMAP test based on the NB model can also be estimated by equation (2.6).

2.3 Simulation Studies

In this section, we evaluate the proposed tests and some existing methods with three simulation studies. For each simulation setting, we simulated 50 independent datasets with each dataset containing 10,000 genes, 2 treatment groups and n replicates for each treatment group where n varied between 2, 3, 5 and 10.

2.3.1 Data Simulation

Simulation A: Poisson Model-Based

. For this simulation, data were simulated from independent Poisson distributions. First, we estimated the distribution of $\pi(\lambda, \delta)$ for the RNA-seq dataset analyzed in Sultan et al. (2008) by fitting a 3-component MGN distribution (see section 2.2.5). Given the estimated parameters $\{(q_k, \alpha_k, \beta_k, \mu_k, \sigma_k) : k = 1, \cdots, 3\}$ (see Table 2.1), we drew λ_g and δ_g from the MGN distribution independently. Then, $p_0 \times 100$ % of the genes were randomly chosen and their δ_g values were set to be zero. Finally, the λ_{gij} was calculated based on equation (2.1), where the normalizing factors S_{ij} for all i and j were set to be 1, and then N_{gij} was generated from the Poisson(λ_{gij}) distribution.

Simulation B: Negative-Binomial Model-Based

. The mean expression level λ_{gij} was generated in the same way as in Simulation A. The dispersion parameters ϕ_g were independently drawn from a Gamma distribution with mean $\alpha/\beta = .85/2$ and variance $\alpha/\beta^2 = .85/2^2$, following the simulations of Hardcastle and Kelly (2010). Then the count N_{gij} was drawn from a NB distribution with mean λ_{gij} and dispersion parameter ϕ_g.

Simulation C: Real Data-Based

. This simulation was based on a large population-based RNA-seq experiment that sequenced 69 lymphoblastoid cell lines (LCL) derived from unrelated Nigerian individuals (Pickrell et. al, 2010). The samples were sequenced at two separate labs (Argonne and Yale) on Illumina Genome Analyzer II instruments, but the two labs generated reads with different lengths. We only selected one lane for each individual from those sequenced at Yale. For each simulation we randomly selected $2n$ out of the 69 individuals and randomly assigned n to one hypothetical treatment group and the remaining n samples to the other hypothetical treatment group. Then, 10,000 genes were randomly selected after excluding those with zero counts across all individuals in both treatments. We expect no differential expression for these genes because the samples were randomly picked from the same population. Then a random sample of $(1-p_0)\times 100\%$ of the selected genes were set to be DE, and their counts in the first and second treatment group were multiplied by $\exp(-\delta_g/2)$ and $\exp(\delta_g/2)$, respectively, where δ_g was drawn from a $N(0,1)$ distribution. The scaled numbers were rounded to the nearest integers. This simulation setting is expected to best mimic the real data because all the counts were originated from real data and no distributional assumptions were imposed.

2.3.2 Simulation Results

2.3.2.1 Testing for Differential Expression

We test the null hypothesis $\delta_g = 0$ for each gene, which will be referred to as testing for differential expression (DE). For each dataset, we estimated the distribution, $\pi(\lambda, \delta)$, by the method described in section 2.2.5, and calculated the AMAP statistics with the estimated distribution. We used the Poisson likelihood for simulation A and the NB likelihood for simulations B and C to calculate the AMAP statistics. Fisher's exact test, edgeR, DESeq and baySeq were also applied to each dataset for comparison. To evaluate the test performance without the influence of different normalization methods, we use the same normalization factors for all tests except Fisher's exact test. Specifically, we set all normalization factors to be 1 for simulations A and B. For simulation C, we set S_{ij} to be the total read count for replicate j of

treatment i before modifying the counts to generate DE genes.

Receiver Operating Characteristic (ROC) curves that plot the true positive rate (TPR) versus the false positive rate (FPR) are shown in Figures 2.1(a) and 2.1(b) for results with $n = 5$. ROC curves for different tests when $n = 2, 3$, and 10 are shown in Figure 2.6. These curves are results of averaging over 50 datasets. We plotted the curves over the FPR values between 0 and 0.1 because the range of small FPR values are of the most practical importance. In addition, we calculated the area under the curve (AUC) for the same range of FPR. The average values and standard errors of the AUC from the 50 simulated datasets are reported in the figure legends. The AUC values presented in all figures are the percentages of 0.1, where 0.1 is the total area for the plotted range of FPR.

For simulations A and B, we also calculated the MAP statistics by equation (2.5) with the true distribution of $\pi(\lambda, \delta)$ used to simulate data. Although this is not available in practice, based on the derivation in section 2.2, the MAP test should provide the highest average power, and hence it is also included for the evaluation of other tests.

Figures 2.1(a) and 2.1(b) shows that the MAP test indeed generated the highest ROC curve and largest AUC among all tests as we expected. We also find that the ROC curves of the AMAP and MAP tests are almost identical for simulation A, and are close for simulation B. The performances of Fisher's exact test, edgeR, DESeq and baySeq are comparable for simulation A, with edgeR and DESeq being slightly better. For simulation B, Fisher's exact test performs much worse than the others, and the baySeq method is the best among edgeR, DESeq and baySeq. In both simulations A and B, the AMAP test significantly outperforms all these tests.

For simulation C, the MAP test is not included in the comparison because we do not know the true distribution $\pi(\lambda, \delta)$. The performance of the AMAP test is superior to that of the other tests. For small values of FPR, the improvement of average power is dramatic. When the FPR is 0.01, the TPR for the AMAP test almost doubles the TPR for edgeR and DESeq (Figure 2.1(a)). Because this simulation setting does not depend on parametric assumptions but is based on real RNA-seq data, the results show that the AMAP test is robust to our model assumptions and can be expected to provide better rankings of genes compared with the other tests when applied to real data.

In Appendix 2.A.4, we present more simulation results with varying sample sizes $n = 2, 3,$ and 10 (Figure 2.6). For simulation A, the performance of AMAP test is indistinguishable from the MAP test for all sample sizes, and both the MAP and AMAP tests are superior than the other tests. For simulation B and C, the AMAP is outperformed when $n = 2$ (Figure 2.6(a)). When $n = 3$, AMAP and baySeq have similar performance that is better than other tests (Figure 2.6(b)). When $n = 10$, AMAP is clearly better than all other tests (Figure 2.6(c)) as shown here when $n = 5$ (Figure 2.1).

We also estimated the FDR for all tests using the AMAP test statistics and equation (2.6) as described in section 2.2.4. The true proportion of false positives among all declared positives at each level of the estimated FDR was plotted in Figures 2.1(c). The estimated FDR levels by our proposed method for all these tests are almost identical to the true values for simulation A. For simulations B and C, the estimated FDR levels are very close to the true values but with slight underestimation. Table 2.1(a) presents the average proportions of false positives (true FDRs) and standard errors for all tests when we control the FDR level at 5% using our proposed method. The true FDRs are between 3.6% and 4.9% for simulations A and C except for one that is 5.5%, and they range from 5.2% to 5.7% for simulation B. For comparison, other widely-used FDR controlling procedures proposed by Benjamini and Hochberg (1995) and Storey and Tibshirani (2003) were also applied to the p-values produced by Fisher's exact test, edgeR and DESeq. The two procedures were not applied to baySeq because baySeq does not provide p-values. In contrast to the good estimation of FDR using the AMAP test statistics, neither of these two procedures controlled the FDR satisfactorily in any of the simulations (Figure 2.7).

Combining the results for the ROC curves and FDR control, we conclude that the proposed AMAP tests generate better rankings of genes in most simulation settings and provide accurate estimation for FDR, and hence provide more reliable lists of DE genes at a desired level of FDR control.

2.3.2.2 Simulation with Outliers

Li and Tibshirani (2011) showed that in real data, there exist outliers that are not well modeled with the negative binomial model. To check the effect of outliers on the performance

of the AMAP test, we modified simulation B (with $n = 5$ and $p_0 = 80\%$) according to the simulation conducted in section 3.2 of Li and Tibshirani (2011). Specifically, after λ_{gij} was simulated, we randomly sampled 1% of these λ_{gij} and set $\lambda_{gij} \leftarrow 10\lambda_{gij}$.

Comparing the results for data with outliers (Figure 2.2) with the results for data without outliers (the middle panel of Figure 2.1(b)), the MAP test assuming no outliers clearly suffer a lot by the introduction of outliers as the AUC drops from 50.8% to 35.4%. The AUCs of all the other tests, edgeR, DESeq, baySeq and Fisher's exact test, are dramatically decreased too. However, the AMAP test is only modestly affected, with AUC dropping from 46.66% to 44.05%. As a consequence, the results show that AMAP test is superior than all the other tests when there exist outliers (Figure 2.2). Again, this show that while the MAP test is only optimal under the model assumptions, the AMAP test is pretty robust and performs well even the model assumptions are violated.

2.3.2.3 Testing for Fold-Changes

Sometimes, biologists want to detect genes whose expression change between treatment groups is large enough (MacCarthy and Smyth, 2009; Covshoff et al., 2008). A common practice is to apply a two-step procedure. The first step is to select a list of DE genes by testing $\delta_g = 0$ while controlling FDR at a certain level. The second step is to select the genes that have large enough fold-changes (FC), such as FC>1.5, among the list of DE genes identified in the first step. The FC can be estimated by the ratios of the mean normalized counts between the two treatments groups. However, the power of such a procedure is not well studied, and oftentimes, the FDR is not estimated for the list of genes detected by this procedure.

Within the framework introduced in section 2.2, we can apply the MAP test and the AMAP test for the null hypothesis $\delta_g \in \Delta_0$ where $\Delta_0 = \{\delta : |\delta| \leq \log(FC)\}$ in order to detect interesting genes with the FC exceeding a pre-determined threshold. We call this testing for FC, which is more general than testing for DE where $\Delta_0 = \{0\}$. Again, the AMAP test statistics and FDR estimation can be calculated by equation (2.5) and formula (2.6), respectively. The ROC curves in Figures 2.3(a) and 2.3(b) show that the AMAP test performs almost identically to the MAP test, and both the MAP and AMAP tests clearly perform better than the two-

step procedure using other tests, especially for the range of small FPR values that is of more practical interest. Figure 2.3(c) shows that the estimated FDR levels by our proposed method for all these tests are almost identical to the true values for simulation A. For simulations B and C, the estimated FDR levels are very close to the true values but with slight underestimation. Table 2.1(b) presents the average proportions of false positives (true FDRs) and standard errors for all tests when we control the FDR level at 5% using the proposed method. The FDR is well controlled for all cases except two where the FDR control is slightly liberal (5.7% and 6%).

Note that we have used the same normalization method for the AMAP, edgeR, DESeq and baySeq tests to obtain all the above simulation results. In real data analysis, we need to estimate the normalization factor (S_{ij}) to adjust for varying sequencing depths and potentially other technical effects across replicates. We compared different normalization methods in the Appendix 2.A.6. When there are symmetric differential expression, the total count method, the third quartile method (Bullard et al., 2010), and the median method by Anders and Huber (2010) perform similarly for AMAP test (Figure 2.8). When the differential expression effect is not symmetric, the total count is worse than the other two normalization methods. In both settings, the AMAP tests with all three normalization methods are superior to DESeq and edgeR test when compared with the default normalization method for associated R packages.

2.4 Real Data Analysis

In this section, we analyze a real RNA-seq dataset published by Li et al. (2010). The dataset can be downloaded from the NCBI short read archive under accession number SRA012297. In this experiment, the maize leaf transcriptome was quantified using Illumina Genome Analyzer II. The dataset includes measurements of transcript abundance of two cell types, bundle sheath and mesophyll, for the tip of maize leaf at a well-defined developmental stage. Each cell type has two biological replicates. In this article, we are interested in comparing the gene expressions between the two cell types.

We assume NB models for the expression counts observed for each gene, and we perform Fisher's exact test, edgeR, DEseq and baySeq tests, and the AMAP test as described in section

2.3.2.1 and 2.3.2.3. We also estimated the FDR levels using formula (2.6) for all these tests. The numbers of detected DE genes at different FDR levels are shown in Figure 2.4(a). Among all testing methods, the AMAP test detected the most DE genes, or equivalently, the estimated FDR level for the AMAP test was the smallest if we declared the same number of positive genes for all applied tests. Moreover, the majority of genes detected by other methods were also identified by the AMAP test. For example, as shown in Figure 2.4(a), when the FDR is controlled at 1%, the AMAP test detected 5537 of the 5891 genes detected by edgeR, and in addition, 3491 genes were detected by the AMAP test but not by edgeR. Note that detecting the most DE genes is not necessarily an indicator of the best method. Follow-up experiments to confirm the extra detected genes will help evaluation of the AMAP method.

We also tested for genes that have expression fold changes exceeding a threshold, FC =1.2 or 1.5, by testing hypotheses $H_0 : |\delta_g| \leq \log(FC)$ with the AMAP test. Not surprisingly, as shown in Figure 2.4(b), when the threshold increases, less genes are detected, and the positive genes rejected at a higher threshold are always detected at a lower threshold.

2.5 Discussion

In this article, we provide a framework for finding the optimal test for RNA-seq data, i.e., the one that maximizes the average power while controlling the FDR. We derive the MAP tests under Poisson and NB models, respectively, and also provide the approximation of the optimal tests, AMAP tests, to be applied in practice. The simulation studies show that the proposed methods perform better than edgeR, DESeq, baySeq and Fisher's exact test. Excluding the proposed methods, baySeq performs better than all other methods. In fact, baySeq can also be viewed as an approximated MAP test where the joint prior distribution is estimated empirically. This helps to explain the near-optimal behavior of the baySeq method.

The results of the AMAP tests are numerically indistinguishable from that of the MAP tests for the Poission model, while the difference between the AMAP and MAP tests is more obvious for the NB model. One reason for this is that there is an extra dispersion parameter in the NB model, and it is challenging to obtain a good estimate or an approximate distribution of this parameter. The estimation of the dispersion parameter for the NB model in the context of

RNA-seq data analysis has drawn attention recently (Robinson and Smyth, 2007; Anders and Huber, 2010; Hardcastle and Kelly, 2010). In Appendix 2.A.7, we compared the performance of the AMAP tests using three different methods to estimate dispersion parameters: the quasi-likelihood (QL) approach, and the edgeR and DESeq approaches proposed by Robinson and Smyth (2007) and Anders and Huber (2010), respectively. We found that when $n = 2$, the QL approach is slightly worse than edgeR and but slightly better than DESeq, while when $n = 3, 5$ and 10, QL method performs better than the other two for simulation B. For simulation C, the QL approach is better than edgeR and DESeq for all sample sizes $n = 2, 3, 5$ and 10. The AMAP tests with estimated dispersion parameters performs worse than the AMAP test with true dispersion parameter and the same estimated $\pi(\lambda, \delta)$. This suggests that better estimation method of the dispersion parameter will likely further improve the performance of the AMAP test.

Although the proposed method is illustrated within the context of RNA-seq data analysis, the AMAP test is also applicable to ChIP-seq data to identify genomic regions of protein occupancy by testing the null hypothesis $\delta_g \in (-\infty, 0]$. Moreover, the framework shown in this article gives a general approach to build optimal tests in multiple hypothesis testing problems.

The R package, named `AMAP.Seq`, is publicly available on `http://www.r-project.org` for implementation of our methods. Users can choose either Poisson or NB distribution to model the counts and specify their own estimates of the normalization factors or dispersion parameters. The computation takes about 45 minutes for a typical RNA-seq dataset with $G = 10,000$ and $n_1 = n_2 = 5$ using a Windows machine with a 3.4GHz CPU and 8GB RAM.

2.6 Acknowledgement

This article is supported in part by the National Science Foundation Grant IOS-0701736.

2.7 APPENDICES

2.A.1 Proof of the Optimality of the MAP Test in Section 2.2.3 to Maximize the Average Power While Controlling the Average Type I Error Rate

In addition to the notations in section 2.2, we denote $f_i(\boldsymbol{X}) = \int_{\mathcal{R}^+}\int_{\Delta_i} f(\boldsymbol{X}|\lambda,\delta)\pi_i(\lambda,\delta)\mathrm{d}\delta\mathrm{d}\lambda$ for $i = 0$ or 1. Since $\int_{\mathcal{X}} f_i(\boldsymbol{X})\mathrm{d}\boldsymbol{X} = 1$, $f_i(\boldsymbol{X})$ defines the PDF of a distribution for $\boldsymbol{X}$ on $\mathcal{X}$. Then, the average power (2.3) is equal to $\int_{\mathcal{X}} \phi(\boldsymbol{X})f_1(\boldsymbol{X})\mathrm{d}\boldsymbol{X}$, and the average type I error rate (2.4) is $\int_{\mathcal{X}} \phi(\boldsymbol{X})f_0(\boldsymbol{X})\mathrm{d}\boldsymbol{X}$. According to the Neyman-Pearson Lemma, the most powerful test that maximizes $\int_{\mathcal{X}} \phi(\boldsymbol{X})f_1(\boldsymbol{X})\mathrm{d}\boldsymbol{X}$ for each fixed level of $\int_{\mathcal{X}} \phi(\boldsymbol{X})f_0(\boldsymbol{X})\mathrm{d}\boldsymbol{X}$ has a rejection region $\mathcal{C} = \{\boldsymbol{X} \in \mathcal{X} : T^*(\boldsymbol{X}_g) = f_0(\boldsymbol{X})/f_1(\boldsymbol{X}) \leq c\}$ for some critical value c.

2.A.2 EM Algorithm to Estimate the MGN Distribution $\pi(\lambda,\delta)$

In section 2.5 in the main manuscript, we approximate the joint distribution $\pi(\lambda,\delta)$ by a K-component MGN distribution defined in equation (7). Suppose that we have determined the number of components K for $\pi(\lambda,\delta)$, or K_0 for $\pi_0(\lambda,\delta)$ and K_1 for $\pi_1(\lambda,\delta)$ with $K = K_0 + K_1$, then the parameters in the MGN distribution are

$$\boldsymbol{\theta} = \{(q_k, \alpha_k, \beta_k, \mu_k, \sigma_k) : k = 1, \cdots, K\}.$$

Using our settings in section 2.5, all parameters except $\{(\mu_k = \delta_0, \sigma_k = 0) : k = 1, \cdots, K_0\}$ in $\boldsymbol{\theta}$ are unknown if the normal components in $\pi_0(\lambda,\delta)$ are degenerate at $\delta = \delta_0$; otherwise, all parameters in $\boldsymbol{\theta}$ are unknown. We introduce a vector $\boldsymbol{Z}_g = (Z_{g1}, \cdots, Z_{gK})$ for gene g, where Z_{gk} is one or zero according to whether (λ_g, δ_g) is from component k of the mixture distribution or not. Assume that the $\boldsymbol{Z}_g$'s are independent samples from a multinomial distribution that consists of K categories with probabilities $\boldsymbol{q} = (q_1, \cdots, q_K)$. The full data from the hierarchical model are

$$(\boldsymbol{X}, \boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{Z}) = \{(\boldsymbol{X}_g, \lambda_g, \delta_g, \boldsymbol{Z}_g) : g = 1, 2, \cdots, G\},$$

where $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_G)$ are observed, and $(\boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{Z})$ are latent variables. Then the non-observable variables $(\boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{Z})$ and unknown hyperparameters in $\boldsymbol{\theta}$ can be estimated via an EM algorithm as follows:

1. *Initialization.* First obtain the estimates of λ_g and δ_g from the Poisson model using their maximum likelihood estimates (MLE). Assuming there are no degenerate normal components in $\pi(\lambda, \delta)$, classify the genes into K groups by clustering the points $(\log \widehat{\lambda}_g, \widehat{\delta}_g)$ via the K-means method. Assuming a degenerate normal distribution in $\pi_0(\lambda, \delta)$, force all $\widehat{\delta}_g$'s that are close to δ_0, say the ones with $|\widehat{\delta}_g - \delta_0| < 0.1$, to be equal to δ_0, then randomly assign these genes into K_0 groups, and cluster all other genes by their values of $(\log \widehat{\lambda}_g, \widehat{\delta}_g)$ into K_1 groups. Then, in both cases, assuming $\widehat{\lambda}_g$ and $\widehat{\delta}_g$ in group k are independently from the gamma distribution $G(\lambda|\alpha_k, \beta_k)$ and (degenerate) normal distribution $N(\delta|\mu_k, \sigma_k)$, respectively, (α_k, β_k) and (μ_k, σ_k), if unknown, can be estimated by their MLEs. The weight q_k for component k can be initialized by the proportion of genes in group k.

2. *E-step.* With the estimated hyperparameters in $\boldsymbol{\theta}$ from the previous step, the expectations of the latent variables can be calculated. We have

$$\mathrm{E}(Z_{gk}|\boldsymbol{X}_g, \boldsymbol{\theta}) = \frac{q_k f(\boldsymbol{X}_g|\alpha_k, \beta_k, \mu_k, \sigma_k)}{\sum_l q_l f(\boldsymbol{X}_g|\alpha_l, \beta_l, \mu_l, \sigma_l)},$$

where $f(\boldsymbol{X}_g|\alpha_k, \beta_k, \mu_k, \sigma_k)$ is the density function of the conditional distribution of $\boldsymbol{X}_g$ given that (λ_g, δ_g) are from component k of $\pi(\lambda, \delta)$:

$$f(\boldsymbol{X}_g|\alpha_k, \beta_k, \mu_k, \sigma_k) = \int\int f(\boldsymbol{X}_g|\lambda, \delta) G(\lambda|\alpha_k, \beta_k) N(\delta|\mu_k, \sigma_k) \mathrm{d}\lambda \mathrm{d}\delta.$$

Here, $f(\boldsymbol{X}_g|\lambda, \delta)$ is the density function for the Poisson model, and it should be replaced throughout by $f(\boldsymbol{X}_g|\lambda, \delta, \widehat{\phi}_g)$ for the NB model after estimating the dispersion parameter ϕ_g. Furthermore, the expectations of λ_g and δ_g are

$$\mathrm{E}(\lambda_g|\boldsymbol{X}_g, \boldsymbol{\theta}) = \sum_k q_k \int\int \lambda f(\lambda, \delta|\boldsymbol{X}_g, \alpha_k, \beta_k, \mu_k, \sigma_k) \mathrm{d}\lambda \mathrm{d}\delta$$

and

$$\mathrm{E}(\delta_g|\boldsymbol{X}_g, \boldsymbol{\theta}) = \sum_k q_k \int\int \delta f(\lambda, \delta|\boldsymbol{X}_g, \alpha_k, \beta_k, \mu_k, \sigma_k) \mathrm{d}\lambda \mathrm{d}\delta,$$

where $f(\lambda, \delta|\boldsymbol{X}_g, \alpha_k, \beta_k, \mu_k, \sigma_k)$ is the conditional distribution of (λ_g, δ_g) given that (λ_g, δ_g) is from component k of $\pi(\lambda, \delta)$:

$$f(\lambda, \delta|\boldsymbol{X}_g, \alpha_k, \beta_k, \mu_k, \sigma_k) = \frac{f(\boldsymbol{X}_g|\lambda, \delta)G(\lambda|\alpha_k, \beta_k)N(\delta|\mu_k, \sigma_k)}{f(\boldsymbol{X}_g|\alpha_k, \beta_k, \mu_k, \sigma_k)}.$$

The computation of integrals can be conducted via the Monte Carlo (MC) approach by drawing random samples of λ from the distribution $G(\cdot|\alpha_k, \beta_k)$ and δ from the distribution $N(\cdot|\mu_k, \sigma_k)$.

3. *M-step.* With the expectation of λ_g, δ_g and $\boldsymbol{Z}_g$ from the previous step, the log-likelihood function, $\ell(\boldsymbol{X}, \boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{Z}|\boldsymbol{\theta})$, for the full data is

$$\sum_k \sum_g Z_{gk} \log\{q_k f(\boldsymbol{X}_g|\lambda_g, \delta_g)G(\lambda_g|\alpha_k, \beta_k)N(\delta_g|\mu_k, \sigma_k)\}.$$

Then the estimates of unknown hyperparameters in $\boldsymbol{\theta}$ can be updated by maximizing this log-likelihood, which is easy because q_k, (α_k, β_k) and (μ_k, σ_k) can be solved separately.

4. *Repeat E- and M-steps until convergence.* We suggest stopping the iteration when the log-likelihood $\ell(\boldsymbol{X}, \boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{Z}|\boldsymbol{\theta})$ changes no more than $G/1000$ from the previous iteration, meaning that the improvement of the log-likelihood is as low as $1/1000$ on average for each gene. According to our experience, the estimates become stable after the first 10 iterations.

2.A.3 Fitting the MGN Distribution to Sultan et al. (2008)'s Data

Sultan et al. (2008) analyzed the transcriptomes from a human embryonic kidney and a B cell line without biological replication. For this RNA-seq dataset, we estimated the model parameters of the MGN distribution defined in equation (7) of the main paper for $K = 1, 2, \cdots, 10$. As shown in Figures 2.5(a)-2.5(c), $K = 1$ provides an unsatisfactory approximation to the distribution of the estimated λ_g and δ_g, while the results from $K = 3$ and $K = 5$ are similar and better than $K = 1$. Figures 2.5(d)-2.5(e) plot the the values of AMAP statistics calculated with estimated distribution of $\pi(\lambda, \delta)$ using different K. The values of the AMAP statistics

for most genes increase when K increases from 1 to 3. However, when K increases from 3 to 5, the changes in the statistics are negligible. We also calculated the Bayesian Information Criterion (BIC) scores for these MGN-Poisson hierarchical models (Figure 2.5(f)), and found $K = 3$ has the smallest BIC value. Hence, $K = 3$ is the number of components we used for estimating $\pi(\lambda, \delta)$. The estimated hyperparameters $\{(q_k, \alpha_k, \beta_k, \mu_k, \sigma_k) : k = 1, 2, 3\}$ are listed in the following table:

k	q_k	α_k	β_k	μ_k	σ_k
1	0.3160810	0.6725020	0.028731200	1.5432904	1.258109
2	0.2232119	0.5027309	0.004668092	-0.6750971	1.528351
3	0.4607071	0.6157050	0.001988262	0.1544223	0.537597

Table 2.1: The hyperparameters for the 3-component MGN distribution estimated by the Expectation-Maximization (EM) for the RNA-seq data from Sultan et al. (2008)

2.A.4 Simulation Results with Different Sample Sizes

For simulation studies A, B and C, we simulated data at different sample sizes. The main manuscript presents results for $n = 5$. Here, we present the ROC curves for testing DE genes for $n = 2, 3$, and 10 (Figure 2.6).

2.A.5 Simulation Results for FDR Control

We performed three simulation studies to evaluate the AMAP tests as described in section 3 of the main paper. We control FDR with the procedures by applying Benjamini and Hochberg (1995)'s or Storey and Tibshirani (2003)'s method to the p-values generated by Fisher's Exact test, edgeR and DESeq for simulation settings with $n = 5$ and $p_0 = 80\%$. Figure 2.7 shows that the FDR control for simulation A is conservative for all three tests. For simulations B and C, the FDR is not controlled for Fisher's exact test. The FDR control is liberal for edgeR and DESeq when true FDR is small, but conservative when true FDR is large (bigger than 0.03). Table 2 present the results of the true FDR when we control the FDR at 5% with our proposed method using the AMAP statistics as described in section 2.4 in the main text. In most cases, the FDR is well controlled. For some cases, the FDR control is slightly liberal.

(a) Test for DE genes

Simulation	A	B	C
Fisher's Exact	4.9 (0.08)	——	3.6 (0.61)
edgeR	4.8 (0.07)	5.2 (0.15)	4.4 (0.15)
DESeq	4.8 (0.07)	5.2 (0.15)	3.9 (0.12)
baySeq	4.8 (0.07)	5.7 (0.11)	4.9 (0.20)
AMAP	4.8 (0.07)	5.3 (0.10)	5.5 (0.29)

(b) Test for FC> 1.5

Simulation	A	B	C
Fisher's Exact	4.7 (0.08)	——	4.2 (0.23)
edgeR	4.7 (0.08)	5.3 (0.14)	4.2 (0.12)
DESeq	4.7 (0.07)	5.1 (0.14)	4.1 (0.20)
baySeq	4.7 (0.08)	6.0 (0.12)	5.2 (0.24)
AMAP	5.0 (0.08)	5.0 (0.10)	5.7 (0.22)

Table 2.2: *FDR Control with the AMAP method.* Data were simulated based on (A) the Poisson model, (B) the NB model and (C) real data with $n = 5$ and $p_0 = 80\%$. We controlled the FDR at 5% for each test with the proposed method using AMAP statistics. The true FDR values were calculated by taking the average of proportions of the false positives among rejections across 50 simulated datasets for each simulation setting. Standard errors are presented in parentheses. All numbers are expressed as percentages. Note that it is impossible to control the FDR to 5% for Fisher's Exact test in simulation B.

2.A.6 Normalization

To evaluate the test performance without the effect of different normalization methods, we have used the same normalization method for the AMAP, edgeR, DESeq and baySeq tests to obtain the all above simulation results and all results presented in the main manuscript. In real data analysis, we need to estimate S_{ij} to adjust for varying sequencing depths and potentially other technical effects across replicates. In this subsection, we compare three ways to estimate S_{ij} and their effect on the AMAP test.

One simple and commonly used way is to estimate S_{ij} by the total number of read for the j-th sample in treatment i (Oshlack et. al, 2010). Because the total read count is largely influenced by a small portion of highly expressed genes, Bullard et al. (2010) proposed a more robust way that estimates S_{ij} by the upper quartile of the non-zero counts within a sample. The DESeq package uses a method that estimates S_{ij} by first calculating the ratio of the gene count in a sample to the geometric mean of this gene across samples and then obtaining the

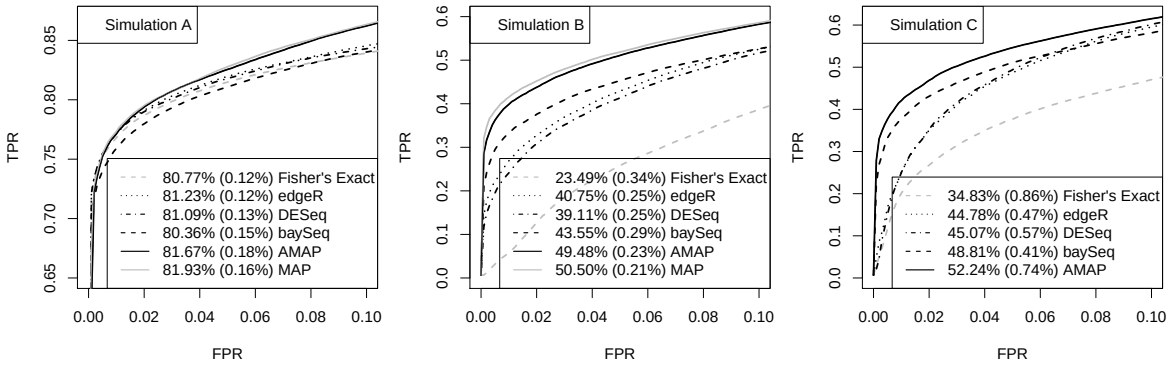
median of those ratios across genes for each sample:

$$S_{ij} = \texttt{median}_g \frac{N_{gij}}{(\prod_{i,j} N_{gij})^{1/\sum_i n_i}}.$$

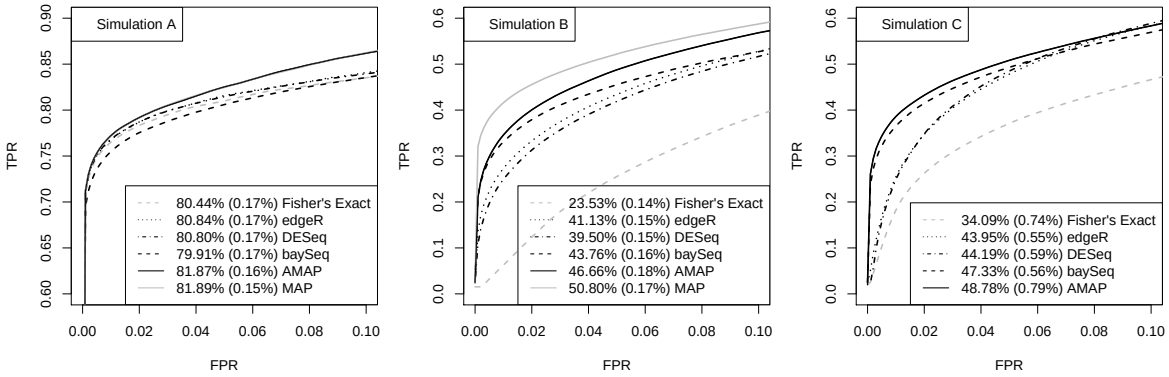
We simulated data according to simulation C with $n = 5$ and $p_0 = 80\%$. To evaluate the normalization methods when the differential expression effects are symmetric or non-symmetric around 0, we simulated δ_g from $N(0, 1)$ and $N(0.2, 1)$ respectively. For each simulated dataset, we applied each of the three normalization methods in the proposed AMAP test. Figure 2.8(a) shows that all three normalization methods applied to the AMAP test perform similarly when there are symmetric differential expression. When the differential expression effect is not symmetric (see Figure 2.8(b)), the median method and the third quartile method perform obviously better than the total count. We compared their performances with the edgeR test using the total count method (the default normalization method of edgeR package), and the DESeq test with the median method of the DESeq package. In both settings, the AMAP tests with all three normalization methods are superior to DESeq and edgeR test.

2.A.7 Simulation Results on Estimation of Dispersion

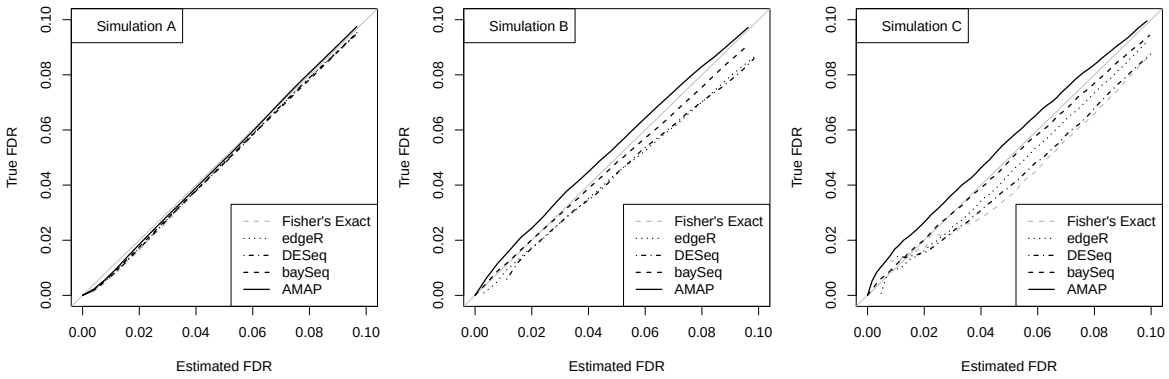
We compared the performance of the AMAP tests using different methods of estimating dispersion parameters. For simulation B, the results show that when $n = 2$, the quasi-likelihood (QL) approach is not as good as the methods by edgeR and DESeq. For $n = 3, 5$, and 10, QL method performs better than the other two methods. As n increases, the ROC curves of the AMAP tests with estimated dispersion parameters approach the ROC curve for AMAP test with the true dispersion parameter (estimated $\pi(\lambda, \delta)$). Also, as n increases, improvement of QL methods over the others becomes more obvious. For simulation C, which is real-data based, the QL method performs best for all sample sizes: $n = 2, 3, 5$, and 10.



(a) ROC curves for testing for DE, $n = 5$, $p_0 = 30\%$



(b) ROC curves for testing for DE, $n = 5$, $p_0 = 80\%$



(c) FDR estimation for testing for DE, $n = 5$, $p_0 = 80\%$

Figure 2.1: *Results from Testing for DE Genes.* Data were simulated based on A: the Poisson model, B: the NB model and C: real data as described in section 2.3.1. (a) The ROC curves that compares the testing powers of different methods when $n = 5$ and $p_0 = 30\%$. For each level of FPR, the TPRs were averaged across the 50 datasets. The percentage annotated for each method is the average AUC, represented as the percentage of the total area 0.1 in the range of FPR<0.1, and the percentage in each set of parentheses is the standard error of the estimated AUC in 50 runs. The grey solid lines for simulations A and B represent the MAP tests that used the true $\pi(\lambda, \delta)$ and ϕ_g from the simulation inputs. (b) The ROC curves that compares the testing powers of different methods when $n = 5$ and $p_0 = 80\%$. (c) The true FDR versus the FDR estimated by equation (2.6).

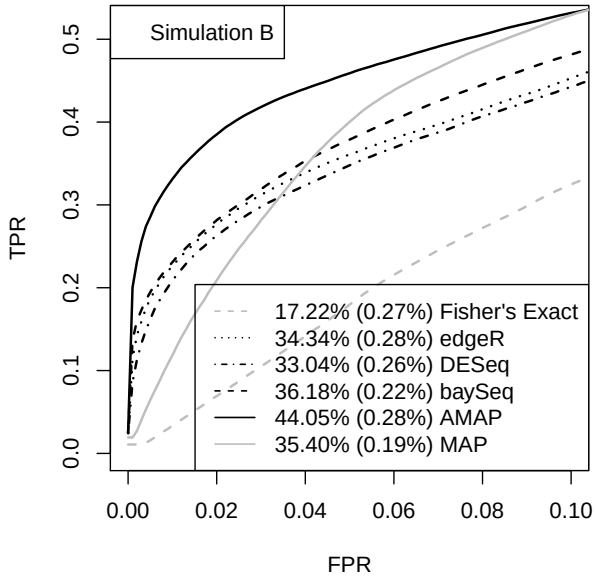
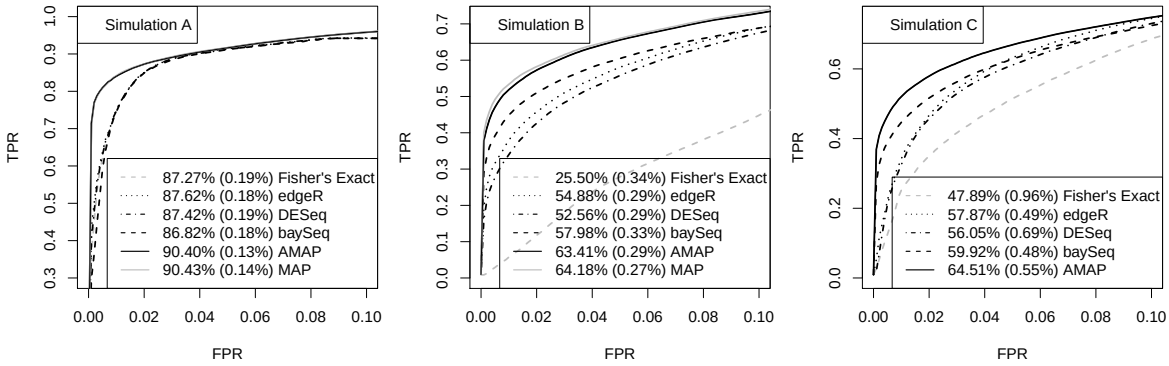
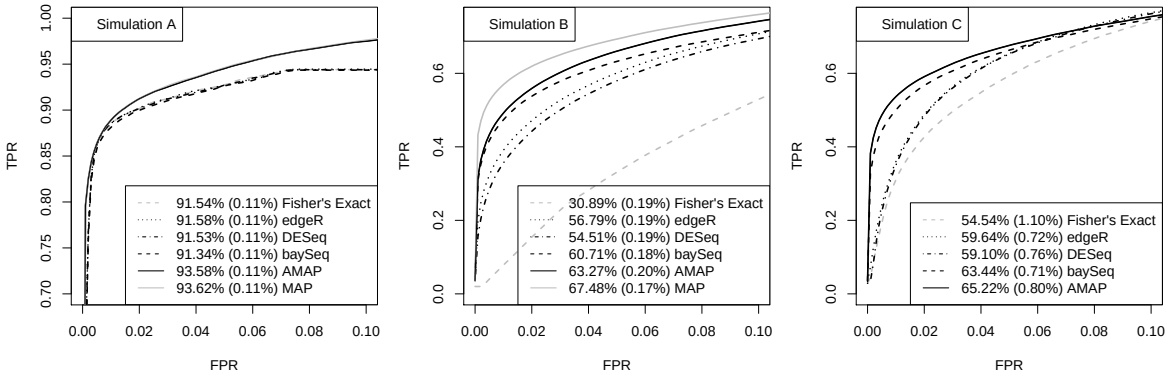


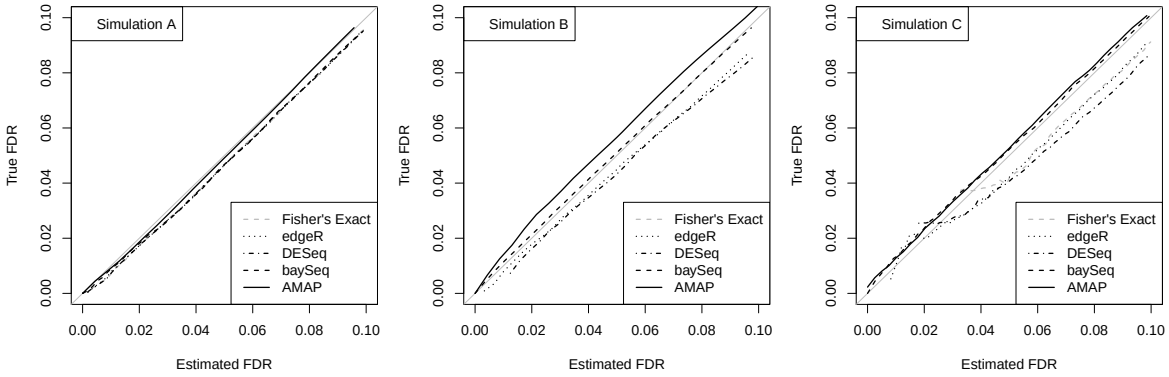
Figure 2.2: *Comparison of the tests in presence of outliers. Data were simulated based on the NB model with* $n = 5$ *and* $p_0 = 80\%$ *as described in section 2.3.2.2. For each level of FPR, the TPRs were averaged across the 50 datasets. The percentage annotated for each method is the average AUC, represented as the percentage of the total area 0.1 in the range of FPR<0.1, and the percentage in each set of parentheses is the standard error of the estimated AUC in 50 runs. The grey solid lines represent the MAP tests that used the true* $\pi(\lambda, \delta)$ *and* ϕ_g *from the simulation inputs before outliers are introduced.*



(a) ROC curves for testing for $FC > 1.5$, $n = 5$, $p_0 = 30\%$

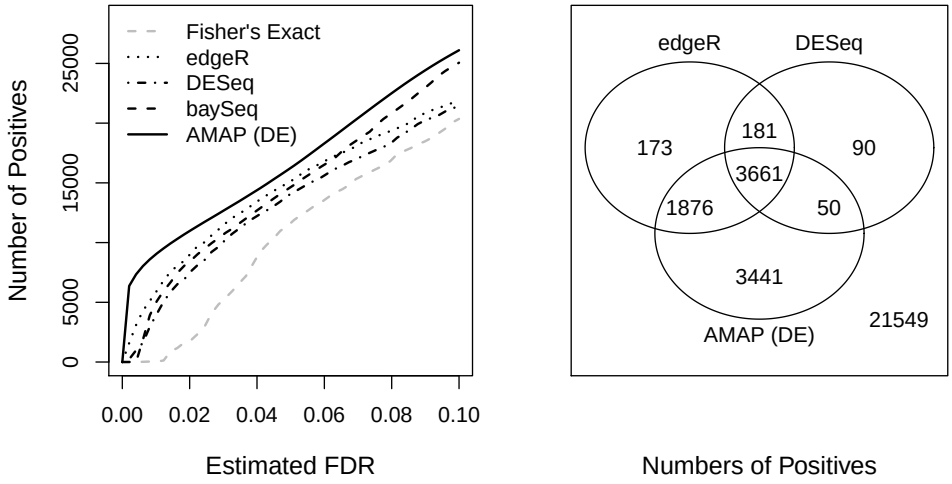


(b) ROC curves for testing for $FC > 1.5$, $n = 5$, $p_0 = 80\%$

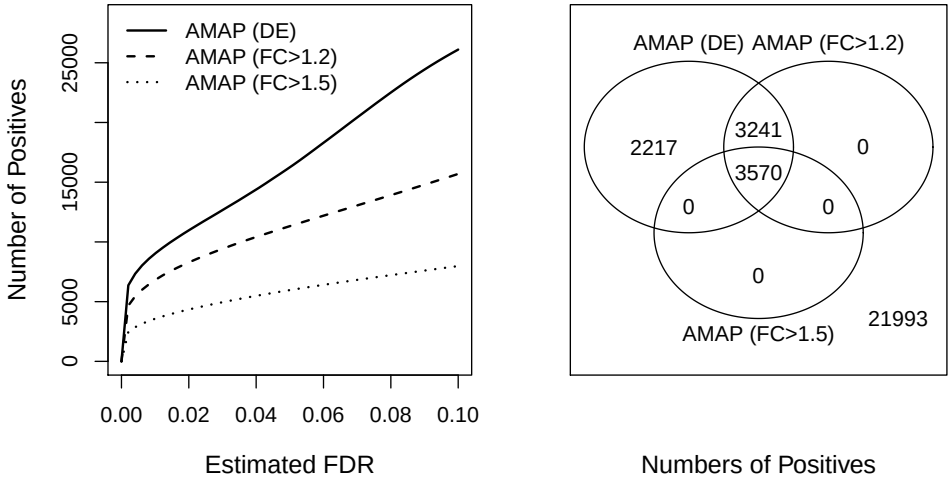


(c) FDR estimation for testing for $FC > 1.5$, $n = 5$, $p_0 = 80\%$

Figure 2.3: *Results from Testing for* $FC > 1.5$. Data were simulated based on A: the Poisson model, B: the NB model and C: real data as described in section 2.3.1. (a) The ROC curves that compares the testing powers of different methods when $n = 5$ and $p_0 = 30\%$. For each level of FPR, the TPRs were averaged across the 50 datasets. The percentage annotated for each method is the average AUC, represented as the percentage of the total area 0.1 in the range of FPR<0.1, and the percentage in each set of parentheses is the standard error of the estimated AUC in 50 runs. The grey solid lines for simulations A and B represent the MAP tests that used the true $\pi(\lambda, \delta)$ and ϕ_g from the simulation inputs. (b) The ROC curves that compares the testing powers of different methods when $n = 5$ and $p_0 = 80\%$. (c) The true FDR versus the FDR estimated by equation (2.6).

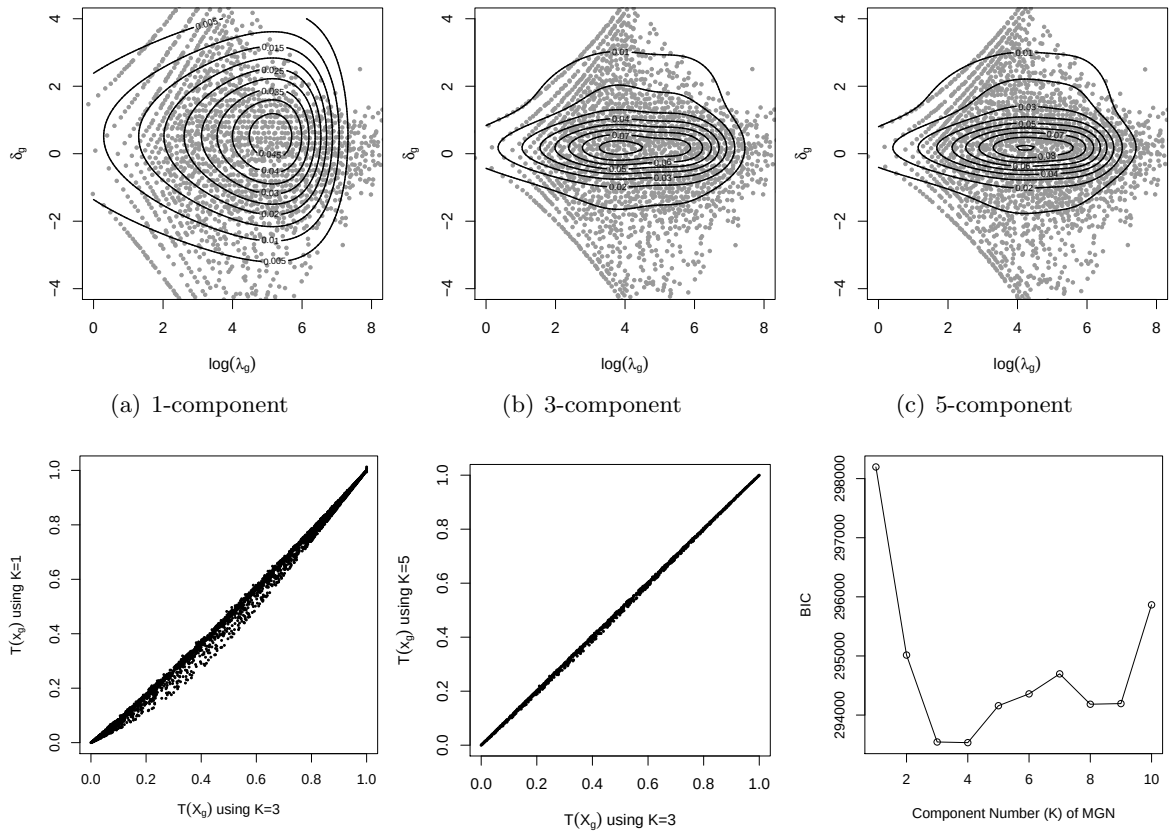


(a) Test for DE genes



(b) Test for genes with FC exceeding 1.0, 1.2, and 1.5

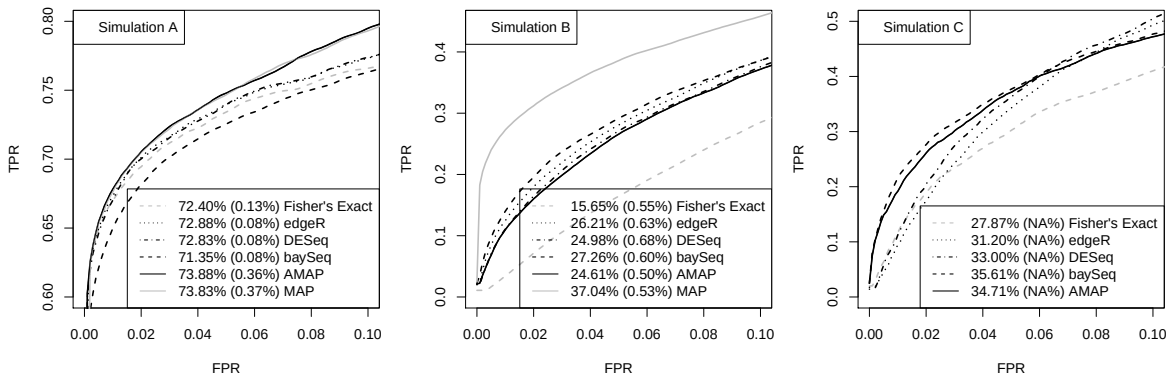
Figure 2.4: *Analysis of Real RNA-seq Data from Li et al. (2010).* (a) Results for different tests of identifying DE genes. The left panel plots the estimated FDR levels when different numbers of genes are declared to be DE. The Venn diagram on the right shows the results of declared DE genes from Fisher's exact test, the edgeR method and the AMAP test when the FDR was controlled at 1% with our proposed method. (b) Results from testing for genes that have FC exceeding thresholds 1.0, 1.2 and 1.5, respectively (testing for FC>1.0 is equivalent to detecting DE genes). The estimated FDR versus the numbers of declared positive genes are compared in the left panel, and the Venn diagram in the right panel shows the numbers of overlapping positive genes declared by the three AMAP tests when the FDR was controlled at 1%.



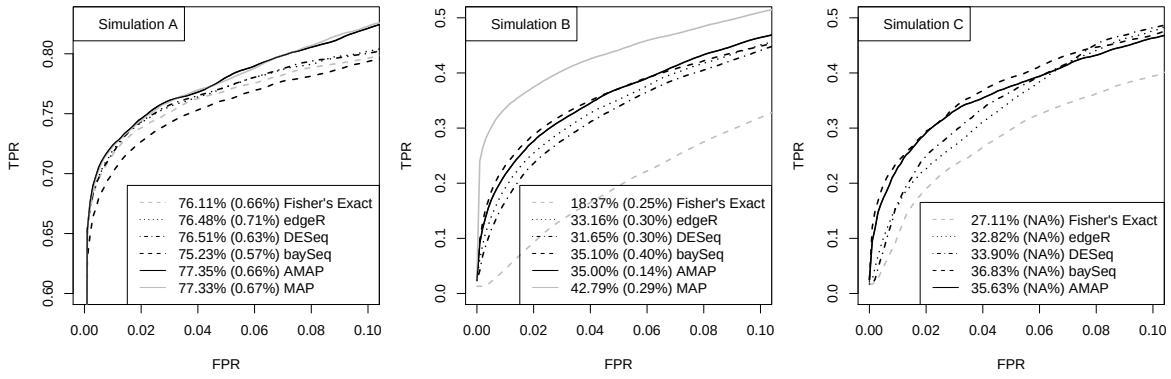
(a) 1-component (b) 3-component (c) 5-component

(d) AMAP Test Statistics (e) AMAP Test Statistics (f) BIC Scores

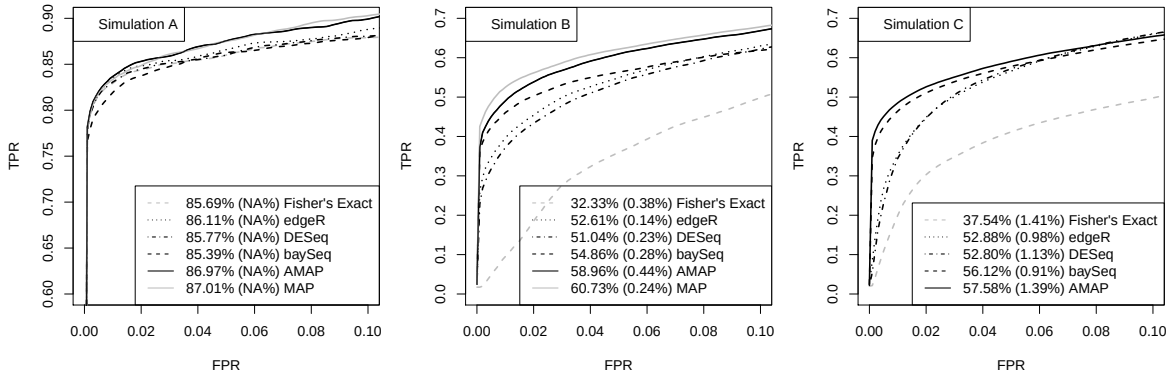
Figure 2.5: *Analysis of the RNA-seq Data from Sultan et al. (2008).* (a-c) Scatterplots of estimated (λ_g, δ_g) and contour plots of $\pi(\lambda, \delta)$ estimated by the EM algorithm using $K = 1, 3$ and 5, respectively. λ_g is on the logarithm scale for easier visualization. (d)-(e) Comparison of the AMAP test statistics $T(\boldsymbol{X}_g)$ resulting from using $K = 1, 3$ and 5. (f) The BIC scores for the MGN-Poisson hierarchical model for $K = 1, 2, \cdots, 10$.



(a) $n = 2$

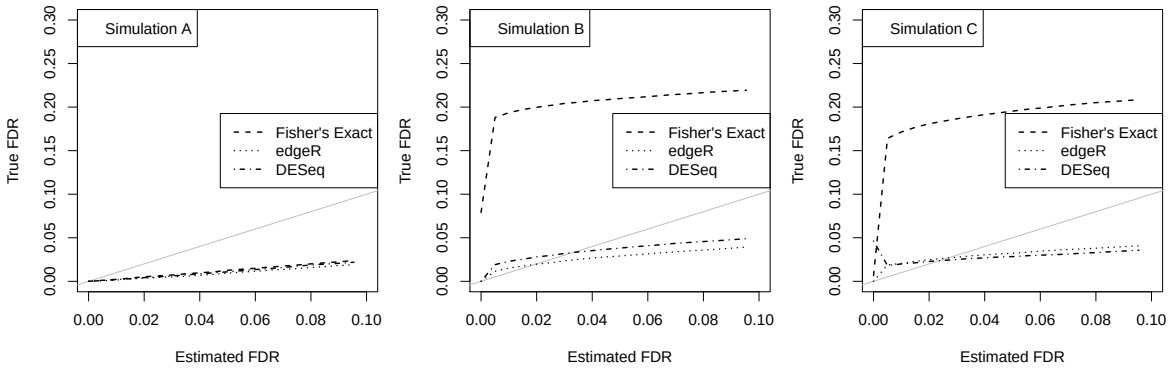


(b) $n = 3$

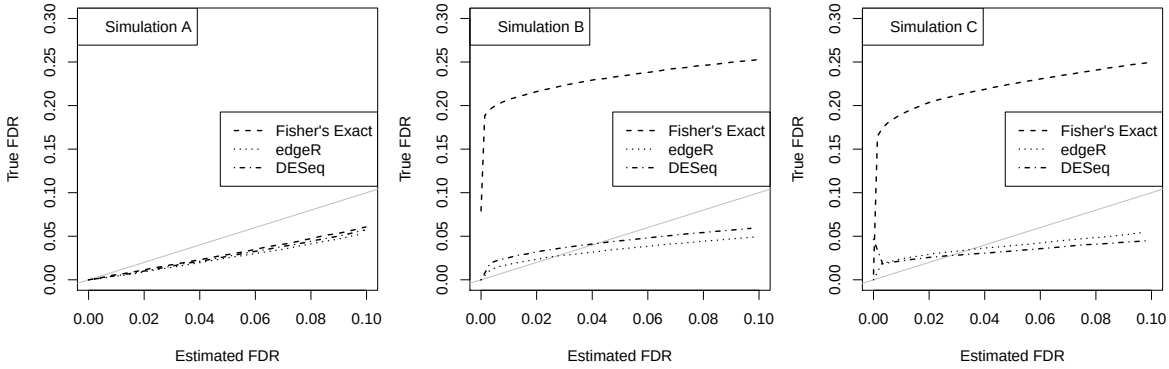


(c) $n = 10$

Figure 2.6: Simulation Results for $n = 2$(a), 3(b),and 10(c) when testing for DE genes. Data were simulated based on A: the Poisson model, B: the NB model and C: real data as described in section 3.1. 50 datasets were simulated for each setting with 10,000 genes in each dataset out of which $p_0 = 80\%$ are non-DE genes. For each level of FPR, the TPRs were averaged across the 50 datasets. The percentage annotated for each method is the average AUC, represented as the percentage of the total area 0.1 in the range of FPR<0.1, and the percentage in each set of parentheses is the standard error of the estimated AUC in 50 runs. The grey solid lines for simulations A and B represent the MAP tests that used the true $\pi(\lambda, \delta)$ and ϕ_g from the simulation inputs.

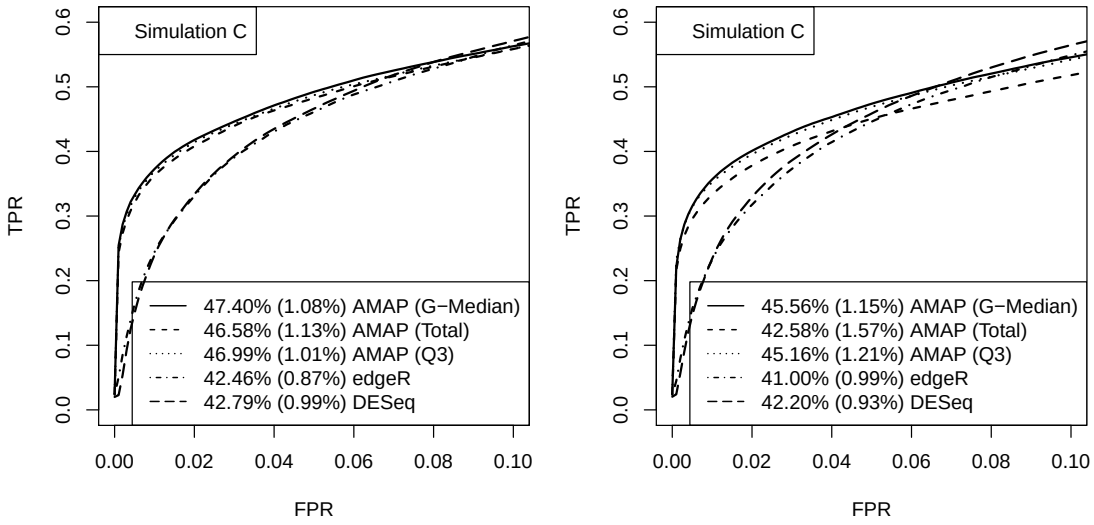


(a) FDR control by Benjamini and Hochberg (1995)'s procedure



(b) FDR control by Storey and Tibshirani (2003)'s procedure

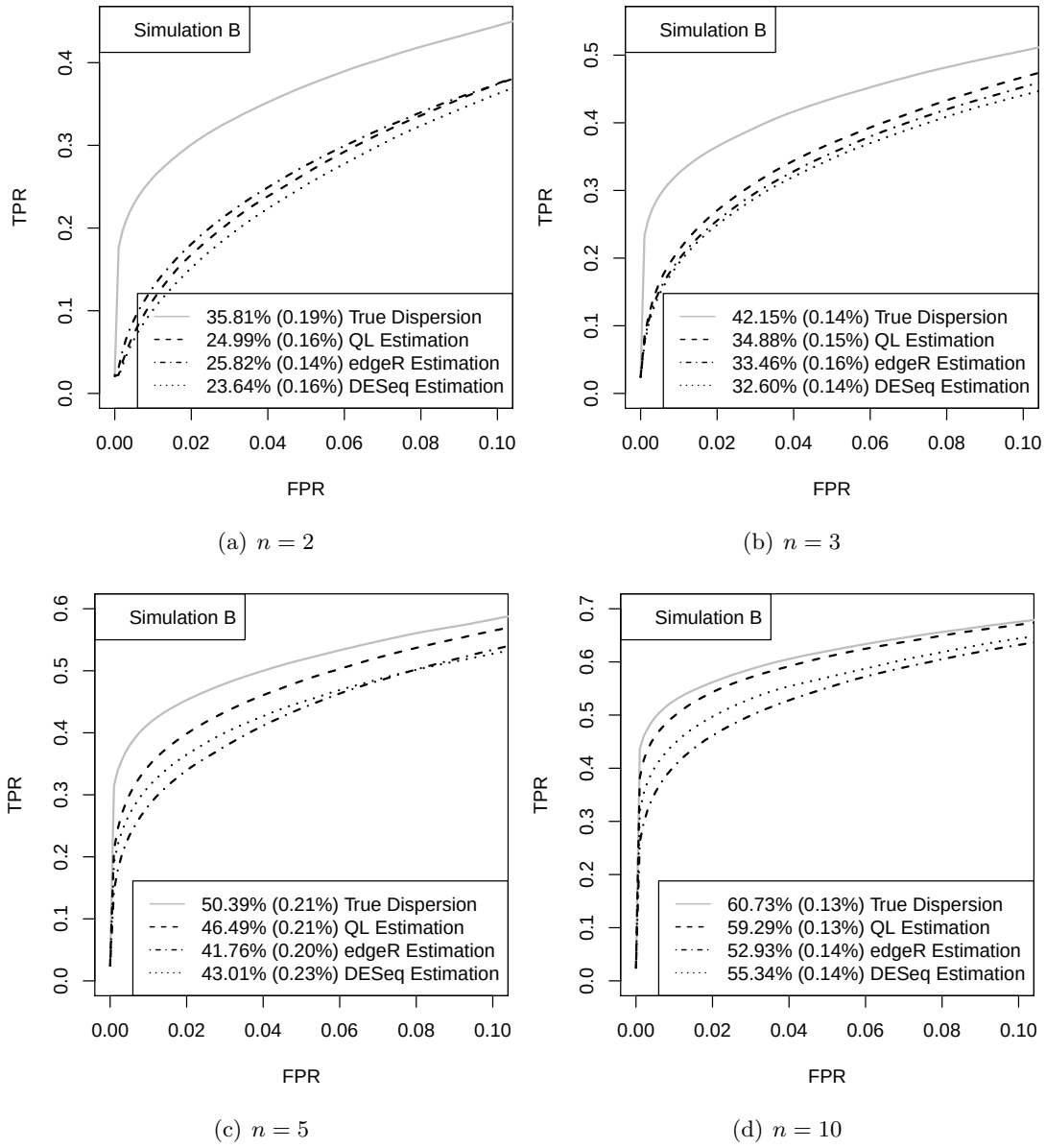
Figure 2.7: *FDR Control When Testing for DE Genes.* Data were simulated based on (A) the Poisson model, (B) the NB model and (C) real data with $n = 5$ and $p_0 = 80\%$. Benjamini and Hochberg (1995)'s and Storey and Tibshirani (2003)'s procedures were applied to the p-values generated by Fisher's exact test, edgeR and DESeq to control the FDR.



(a) Symmetric Distribution of δ_g

(b) Non-Symmetric Distribution of δ_g

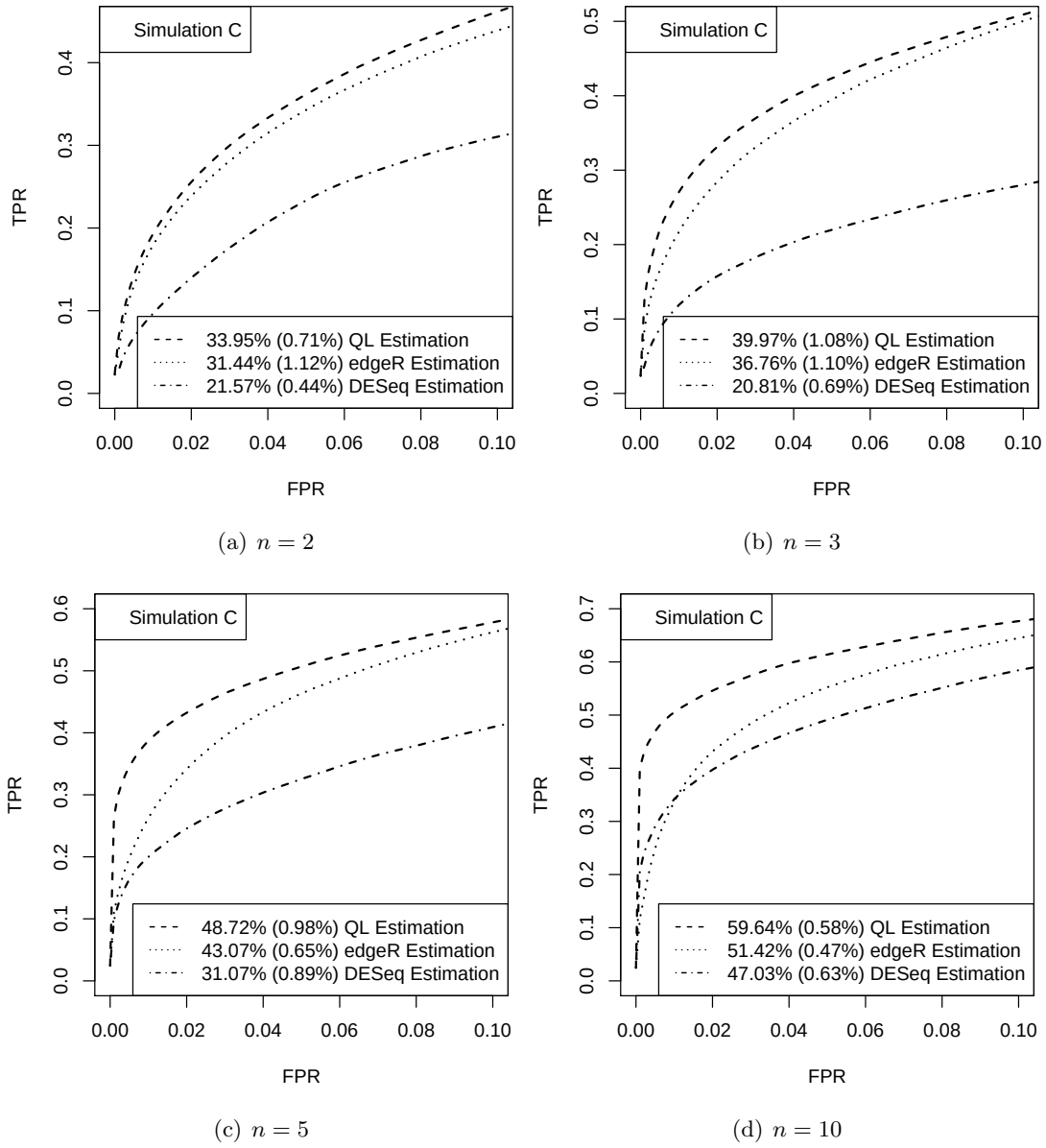
Figure 2.8: *Different normalization methods applied to the AMAP test.* (a) Symmetric Distribution of δ_g simulated from $N(0,1)$. (b)Non-Symmetric Distribution of δ_g simulated from $N(0.2,1)$. Data were simulated based on real data as described in section 3.1. 50 datasets were simulated for each setting with 10,000 genes in each dataset out of which $p_0 = 80\%$ are non-DE genes. For each level of FPR, the TPRs were averaged across the 50 datasets. The percentage annotated for each method is the average AUC, represented as the percentage of the total area 0.1 in the range of FPR<0.1, and the percentage in each set of parentheses is the standard error of the estimated AUC in 50 runs.



(a) $n = 2$ (b) $n = 3$

(c) $n = 5$ (d) $n = 10$

Figure 2.9: *Check Dispersion Estimation for Simulation B.* Data were simulated based on simulation B as described in section 3.1. 50 datasets were simulated for each setting with 10,000 genes in each dataset out of which $p_0 = 80\%$ are non-DE genes. We calculated the AMAP statistics using the true dispersion parameter values (True Dispersion) and three ways to estimate dispersion parameters (QL, edgeR, and DESeq). For each level of FPR, the TPRs were averaged across the 50 datasets. The percentage annotated for each method is the average AUC, represented as the percentage of the total area 0.1 in the range of FPR<0.1, and the percentage in each set of parentheses is the standard error of the estimated AUC in 50 runs.



(a) $n = 2$ (b) $n = 3$

(c) $n = 5$ (d) $n = 10$

Figure 2.10: *Check Dispersion Estimation for Simulation C.* Data were simulated based on simulation C as described in section 3.1. 50 datasets were simulated for each setting with 10,000 genes in each dataset out of which $p_0 = 80\%$ are non-DE genes. We calculated the AMAP statistics using three ways to estimate dispersion parameters (QL, edgeR, and DESeq). For each level of FPR, the TPRs were averaged across the 50 datasets. The percentage annotated for each method is the average AUC, represented as the percentage of the total area 0.1 in the range of FPR<0.1, and the percentage in each set of parentheses is the standard error of the estimated AUC in 50 runs.

CHAPTER 3. Statistical Analysis of Alternative Splicing Events

Yaqing Si and Peng Liu

Iowa State University, Snedecor Hall, Ames, IA 50011, USA.

Abstract

For eukaryotic cells, alternative splicing (AS) is very common in the transcription of a gene, and studying AS events is important to understand gene functions. The recent RNA-seq technology has created great opportunities by offering never before resolution of the exon-level expressions. Though some techniques to detect differential gene expressions can be used in comparing exon coverages, efficient tools specified to analyze exon expressions are still very limited. Some methods such like `DEXSeq` *(Anders et al., 2012) have been developed to test for differential exon usages. However, the detection power of these tests has not been evaluated very well. Moreover, some other interesting AS patterns, such as exon-skipping and switch-like patterns, still need more investigation. We generalize the approximated most average-powerful (AMAP) test from our previous research on testing gene expression data to studying AS patterns. A nonparametric algorithm to estimate the distribution of exon usages is proposed, which provides more flexibility for fitting the data as well as higher efficiency for computation. Our methods is compared with previous methods in a real data-based simulation study and is shown to be much more powerful.*

Key words: Alternative Splicing, Isoforms, Exon Usages, Differential Expressions, Fold Changes, Switch-Like Patterns, Empirical Bayes, Most Average-Powerful.

3.1 Introduction

For eukaryotic cells, it is common that a gene has several protein-coding regions called *exons*, and the exons of a gene are reconnected in multiple ways during RNA splicing. The resulting different mRNAs may be translated into different protein isoforms (see Figure 1.2 for the illustration). This process is called *alternative splicing* (AS). AS affects message stability and translation efficiency, and increases protein diversity (Black, 2003; Stamm et al., 2005). AS in particular is known to affect more than half of all human genes, and has been proposed as a primary driver of the evolution of phenotypic complexity in mammals (Lander et al., 2001; Johnson et al., 2003). So studying AS events has been an important question for scientists.

Previous studies of AS based on exon-arrays have achieved many findings. However, using a hybridization-based approach, microarray is constrained in its ability to distinguish closely related mRNA isoforms (Wang et al., 2008). The recent next-generation sequencing (NGS) technology has been developed to use direct sequencing to measure the coverages of mRNA across the genome, and the resulting RNA-seq data provide digital signals of exon expressions in form of the numbers of mRNA fragments, the so-called *reads*, that are aligned to the coding regions. Then the exon coverages can be used to investigate the abundance of distinct isoforms. Compared with microarray's hybridization approach, the NGS technology creates a unique opportunity by offering never before resolution of the transcriptome, hence is gradually overtaking the former as the mainstream in studies of AS.

Many current researches of RNA-seq data focus on gene-level and transcript-level expressions, and some of these methods may be applied to studying exon coverages. For example, the frequently-used R packages, `edgeR` (Robinson and Smyth, 2008), `DESeq` (Anders and Huber, 2010) and `baySeq` (Hardcastle and Kelly, 2010) that are mainly designed to detect differentially expressed (DE) genes can also be applied to testing for DE exons. These methods treat exons of the same gene in the same way as exons from different genes, i.e., they ignore the relationship of among exons. However, we expect that exons of the same gene are correlated. For example, if a gene is highly expressed, then we expect its exons exhibit high coverage in general. Methods that incorporate such information are still very limited. Hence, there is increasing need for

robust and adaptable statistical/bioinformatics methods to interrogate the NGS data for AS analysis.

Recently, Anders et al. (2012) proposed the `DEXSeq` method to test for differential exon usages. For each gene, this method fits one generalized linear model (GLM) that includes the gene, treatment and exon effects, and calculates the p-value for each exon from the χ^2 likelihood-ratio test based on negative-binomial (NB) distribution. `DEXSeq` uses relationship between the exons of the same gene, and is applicable to multi-treatment comparison. However, due to the asymptotic properties of the χ^2 likelihood-ratio test and the usually small number of replicates in RNA-seq experiments, this method may introduce many false positives (see Figure 3.4), and especially lacks capacity to handle low count exons. Another method, MATS, that is proposed by Shen et al. (2012) provides a Bayesian statistical framework to evaluate the statistical significance that the absolute difference in exon inclusion levels between two treatments exceeds any user-defined threshold. A major advantage of MATS is that it allows flexible hypothesis testing of various AS patterns, for example, testing for user-defined threshold of fold changes (FC), or detecting the extreme 'switch-like' patterns. However, one needs to transform the count data to continuous data to apply this method, and Shen et al. (2012) has reported that the FDR control for MATS is not accurate.

We developed the approximated most average-powerful (AMAP) test to compare gene expressions in Chapter 2 and showed its superior performance. Here, we will generalized the AMAP test to exon-level RNA-seq data with focus on comparing exon usages in a two-treatment experiment. We aim to find a powerful testing procedure while controlling FDR, provide flexible hypotheses choice for detecting different AS patterns, give a good FDR estimation approach, and in addition, improve the efficiency of computing the AMAP test statistics.

This chapter will be organized as follows: section 3.2 models the exon coverages with discrete probability models including Poisson and NB distributions, and incorporates the parameterization of exon usages with the normalization step using varies factors; section 3.3 formulates the hypotheses to test, and gives examples of hypotheses for three interesting AS patterns; section 3.4 builds the AMAP test based on an empirical Bayes (EB) framework; section 3.5 introduces an non-parametric method and an efficient computing algorithm to estimate the distribution

of exon usages; section 3.6 presents a simulation study based on real data, and our proposed AMAP test is compared with other methods with respect of their testing powers; at last in section 3.7, a real RNA-seq data set from rice is analyzed.

3.2 Model

Suppose that the experiment has J_i replicates in treatment group i ($i = 1, 2$), the reads are mapped to G genes, and there are E_g exons for gene g ($g = 1, 2, \cdots, G$). Let the number of reads mapped to the e-th exon of gene g in replicate j of treatment i be denoted by N_{geij} for $j = 1, \cdots, J_i; e = 1, \cdots, E_g$. We view N_{geij} as a random variable with expectation

$$\mathrm{E}(N_{geij}) = \mu_{geij}. \tag{3.1}$$

The mean μ_{geij} depends on several factors. Firstly, one factor is the sequencing depth, or library size of the sample (i, j) that can serve as a between-sample normalization factor S_{ij}. Similar to Chapter 2 where the gene level RNA-seq data are studied, there are several ways to estimate S_{ij}, for example, by the total number of mappable reads (Mortazavi et al., 2008). Other normalization methods such as the 75th percentile (Q3) (Bullard et al., 2010) can also be applied. Secondly, the exon coverages also depend on the gene expression level, and a gene with high expression level is expected to have high exon coverages in general. Let M_{gij} be the *reads per kilobase of exon per million reads* (RPKM) that is used to measure the gene expression (Mortazavi et al., 2008). We assume that μ_{geij} is proportional to M_{gij}. A similar assumption has been made to normalize the intron reads in a RNA-seq study (Wang et al., 2011). The third factor that impacts μ_{geij} is the length of the exon, L_{ge}, because more reads are expected to be mapped to longer exons. To summarize these factors, we have

$$\mu_{geij} = S_{ij} M_{gij} L_{ge} \mu_{gei}, \tag{3.2}$$

where μ_{gei} is expression of the exon in treatment i after normalization.

To interpret μ_{gei} in the equation, suppose that S_{ij} is the total number of mapped reads in the treatment, the read count and length for gene g are N_{gij} and L_g, respectively. If N_{gij} and L_g for the gene are not given, they can be approximated by $\sum_e N_{geij}$ and $\sum_e L_{ge}$, respectively,

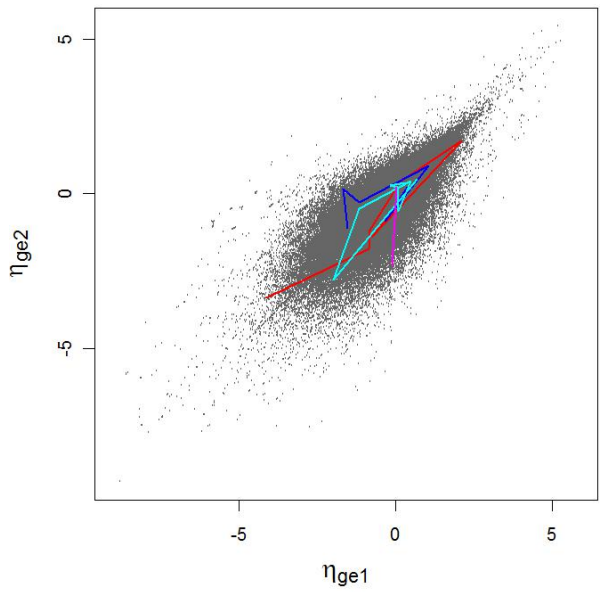


Figure 3.1: *Exon Effects*. Estimate μ_{gei} by $\widehat{\mu}_{gei} = \frac{\sum_j N_{geij}}{\sum_j S_{ij} M_{gij} L_{ge}}$ by method of moments estimation. The points linked by lines of the same color are for a set of exons from the same gene. Transformation $\eta_{gei} = \log(10^9 \widehat{\mu}_{gei})$ is used for better visualization.

using the exon data. Then we have $M_{gij} = \frac{N_{gij} \times 10^9}{S_{ij} L_g}$ by the definition of RPKM . From equation (3.1) and (3.2), we get $\mathrm{E}(N_{geij}) = N_{gij} \times \frac{10^9 L_{ge}}{L_g} \mu_{gei}$ for fixed N_{gij}. Hence $\frac{10^9 L_{ge}}{L_g} \mu_{gei}$ represents the expected fraction of reads mapped to exon e among all reads mapped to gene g, and we call μ_{gei} the *normalized usage* (or simply *usage* without confusion) of the exon coverage with respect to the coverage of the whole gene, where the normalization scalar $\frac{10^9 L_{ge}}{L_g}$ is proportional to the ratio of exon length to the gene length. Considering only treatment i and gene g, μ_{gei} will be be a constant (10^{-9}) for all exons when there are no alternative splicing, and would vary across the exons otherwise. Figure 3.1 plots the point estimates of these μ_{gei} (up to a common scalar 10^9 and the log-transformation) from two treatments, we can see that many exon usages deviate from the diagonal of the scatter plot, which indicates differential alternative splicing between two treatments. So we are interested in comparing each pair of μ_{ge1} and μ_{ge2} for each combination of exon and gene.

Given the expectation μ_{geij}, Vardhanabhuti et al. (2011) assumed a Poisson distribution for N_{geij}. However, as pointed out by several previous studies for RNA-seq data, the Poisson model for RNA-seq data usually suffers the so-called over-dispersion problem (Anders and

Huber, 2010; Robinson and Smyth, 2007), which means the variation of the counts are larger than their means when there are more than one biological replicate in each treatment. Anders et al. (2012) also noticed this limitation of the Poisson assumption, and proposed to overcome it by assuming a negative-binomial (NB) distribution. We adopt the NB assumption and model N_{geij} by:

$$N_{geij} \sim NB(\mu_{geij}, \phi_{ge}), \tag{3.3}$$

where the mean μ_{geij} and dispersion parameter ϕ_{ge} are determined by $\mathrm{E}(N_{geij}) = \mu_{geij}$ as in equation (3.1) and (3.2) and $\mathrm{Var}(N_{geij}) = \mu_{geij} + \phi_{ge}\mu_{geij}^2$.

Among the unknown parameters $\boldsymbol{\mu}_{ge} = (\mu_{ge1}, \mu_{ge2})$ and ϕ_{ge} for in the NB model, we are more interested in the usages $\boldsymbol{\mu}_{ge}$ for the exon. If ϕ_{ge} is known and given the data $\boldsymbol{N}_{ge} = \{N_{geij}\}$, the probability mass function (p.m.f.) for the NB distribution of the data can be written as

$$\boldsymbol{N}_{ge} \sim f(\boldsymbol{N}_{ge}|\boldsymbol{\mu}_{ge}) \tag{3.4}$$

In practice, we need to estimate ϕ_{ge}. Similar to the methods of estimating dispersions for gene expression data as discussed in section 2.2.6 of Chapter 2, we can do the estimation by Quasi-Likelihood (QL) method (Nelder, 2000; Robinson and Smyth, 2008). Anders et al. (2012) also proposed a method based on the work of Cox and Reid (1987). Note that we assume different dispersion parameters in model (3.3) for different exons but ϕ_{ge} are the same across treatment groups for each exon for simplicity. Anders et al. (2012) allow the dispersion parameters to differ between the two treatment groups for each exon. It is straightforward to adapt our method when the dispersion parameters differ across treatments.

3.3 Hypotheses

There are many interesting AS patterns under study. For two-treatment experiments concerning with exon usages, most patterns are usually investigated by multiple-testing procedures that test hypotheses about $\boldsymbol{\mu}_{ge}$. Explicitly, the hypotheses to test for exon e of gene g are

$$H_0^{(ge)} : \boldsymbol{\mu}_{ge} \in \Delta_0^{(ge)} \text{ vs } H_1^{(ge)} : \boldsymbol{\mu}_{ge} \in \Delta_1^{(ge)}, \tag{3.5}$$

where $\Delta_0^{(ge)}(\Delta_1^{(ge)})$ is the null (alternative) space, and $\Delta_0^{(ge)} \cup \Delta_1^{(ge)} = \mathcal{R}^{+2}$. The hypotheses are general enough to handle various AS patterns by specifying proper $\Delta_0^{(ge)}$ or $\Delta_1^{(ge)}$ correspondingly. In this work, we are going to study three interesting patterns of AS as illustrated next. However, the hypotheses (3.5) and our proposed methods in section 3.4 are not limited to the three examples:

3.3.1 Test for Inclusion-Skipping

Though the NGS technology generates data with significantly lower noise compared with microarray, it is not noise-free and it is possible that some low counts are not real signals of expression. Hence, a filtering step is often taken place to exclude the low count exon regions where the mapped reads are actually noises. In addition, maybe more importantly, since many isoforms do not use the all exons, but alternatively select exons and splice sites during RNA splicing, it is common that some exons are skipped without expression, hence finding these skipped exons are helpful to identify distinct isoforms.

Suppose that the RNA-seq data have constant noise level along the genome, then the expected number of mapped reads from the non-coding regions and non-expressed exon regions will depend on the sequencing depth and the length of the region. Therefore, we expect the skipped exons have $\mathrm{E}(N_{geij}) \leq S_{ij} L_{ge} b_{ij}$ reads, where b_{ij} is a constant that represents the noise level of the background in sample (ij) (see Appendix 3.A.1 for details of method to estimate b_{ij}). Then along with equation (3.2), we get $\sum_j S_{ij} M_{gij} L_{ge} \mu_{gei} \leq \sum_j S_{ij} L_{ge} b_{ij}$, so the cutoff for μ_{gei} can be chosen at

$$b_{gi} := \frac{\sum_j S_{ij} b_{ij}}{\sum_j S_{ij} M_{gij}}. \tag{3.6}$$

When $\mu_{gei} \leq b_{gi}$, we say the exon is skipped or not expressed in the treatment. Most inclusion-skipping AS patterns can be studied by comparing μ_{gei} with proper b_{gi}. For example, to test for expressed exon in at least one of the two treatments, we set

$$\Delta_1^{(ge)} = \{(x, y) : x > b_{g1} \text{ or } x > b_{g2}\} \tag{3.7}$$

3.3.2 Test for Switch-Like Pattern

The switch-like differential AS pattern means an exon is predominately included in the transcripts in one treatment but predominately skipped in another (Shen et al., 2012). So the switch-like pattern is a special inclusion-skipping event. Biologically, a switch-like pattern strongly indicates structural and functional changes for the isoforms during AS, hence is of particular interest (Wang et al., 2008; Xing and Lee, 2005). To detect this kind of events, we set the alternative set as

$$\Delta_1^{(ge)} = \{(x, y) : 0 \leq x \leq b_{g1} \ \& \ y > b_{g2}\} \cup \{(x, y) : x > b_{g1} \ \& \ 0 \leq y \leq b_{g2}\} \tag{3.8}$$

3.3.3 Test for Fold Changes

Distinct isoforms for different treatments are the products of inclusion-skipping events. More frequently, the products of AS are combinations of different isoforms for each treatment group. Hence changes in AS are often slight shifts (by as few as several percent) in the relative abundance of multiple mRNA isoforms (Shen et al., 2012). So it is often desirable for many biologists who are more interested in the differences of exon usages that exceeds a user-defined cutoff, and we can set the alternative set using

$$\Delta_1^{(ge)} = \{(x, y) : |\log(x/y)| > c\} \cap \{(x, y) : x > b_{ge1} \text{ or } y > b_{ge2}\} \tag{3.9}$$

where c is a user-defined threshold, say $c = \log(1.2)$ if fold-changes (FC) exceeding 1.2 are going to be detected. Criterion $x > b_{ge1}$ & $y > b_{ge2}$ are added to exclude non-expressed exons from the positive list.

Compared with some methods that only test for the equality of exon usages such as `DEXSeq`, our settings in 3.9 provides convenience to biologists for selecting exons that reach any magnitude of FC. And also, as pointed by Shen et al. (2012), the random sampling noise in the RNA-seq data may cause a minor shift in the estimated ratio between the exon usages, and the shift will introduce false positives if assuming the exon usages from the null are exactly equal, hence testing for FC can improve the robustness against the inaccurate estimation.

3.4 The AMAP Test

In this section, we are going to build a multiple testing procedure to test hypotheses (3.5) for all exons, where $\Delta_1^{(ge)}$ could be in forms of (3.7), (3.8) or (3.9). Note that there are usually tens of thousands of exons to be tested simultaneously, and we care about the overall performance of the test instead of its behavior on any individual exon. It has been widely noticed that borrowing information across different observations can be useful to improve the overall performance for multiple testings, and empirical Bayes (EB) has been proved as a powerful framework under which proper testing procedure can be derived. For example, in developing the so-called baySeq method to study differential gene expressions, Hardcastle and Kelly (2010) assume all gene expressions are from a prior distribution and calculated the posterior probability of the null model for each gene. In another study, Wang et al. (2011) assume the expressions of the introns are from a shared Gamma distribution, and the resulting testing method by borrowing information has good performance to detect intron retention events.

In Chapter 2, we have adopted the EB framework to develop the *most average-powerful* (MAP) test to compare gene expressions between two treatments from RNA-seq data. Under reasonable assumption, the MAP test has been shown to be optimal by maximizing the average power while controlling the false discovery rate (FDR). The MAP test can be adapted to exon data. We assume that all exon usages $\{\boldsymbol{\mu}_{ge}\}$ share a distribution:

$$\boldsymbol{\mu}_{ge} \sim \pi(\boldsymbol{\mu}), \tag{3.10}$$

where $\pi(\boldsymbol{\mu})$ is the probability density function (p.d.f.) of the shared 2-dimentional distribution. Under this assumption and using arguments similar to that in section 2.2.3 of Chapter 2, we can derive the MAP test statistic in form of

$$T(\boldsymbol{N}_{ge}) = \frac{\int_{\Delta_0^{(ge)}} f(\boldsymbol{N}_{ge}|\boldsymbol{\mu})\pi(\boldsymbol{\mu})\mathrm{d}\boldsymbol{\mu}}{\int_{\mathcal{R}^{+2}} f(\boldsymbol{N}_{ge}|\boldsymbol{\mu})\pi(\boldsymbol{\mu})\mathrm{d}\boldsymbol{\mu}}, \tag{3.11}$$

In practice, we need to estimate $\pi(\boldsymbol{\mu})$ from the data, and that will result in an approximated MAP (AMAP) test. We reject the null hypothesis in (3.5) if the AMAP test statistic $\widehat{T}(\boldsymbol{N}_{ge})$

is small, and the cutoff can be chosen by controlling FDR, using formula (2.6) in section 2.2.4 of Chapter 2.

3.5 Computation

To calculate the AMAP test statistic by equation (3.11), it is important to get a good estimator of the prior distribution $\pi(\boldsymbol{\mu})$. In section 2.2.5 of Chapter 2, we proposed a data-driven method that, if applied here, models $\pi(\boldsymbol{\mu})$ using a K-component mixture distribution. The mixture distribution provides convenience to fit different dataset by flexibly choosing the number of components and the hyper-parameters in each component. Compared with the method by Hardcastle and Kelly (2010) which approximated the prior distribution by the empirical distribution of the maximum likelihood estimate (MLE) across all genes, the K-component mixture distribution employed the shrinkage technique that does the estimation by borrowing information across different genes, which has been shown to be able to improve the overall performance of the multiple testing procedure (see section 2.3 of Chapter 2). However, one needs to decide a proper component number, K, for the mixture distribution, and the expectation-maximization (EM) algorithm to estimate hyper-parameters often requires intensive computation.

In this section, we will introduce another method that estimates $\pi(\boldsymbol{\mu})$ with a non-parametric method. Before doing this, we notice that the distribution of exon counts is often highly skewed to the left. So we first apply log transformation $\eta_{gei} = \log(\mu_{gei})$ to exon usage $\boldsymbol{\mu}_{ge}$, then equivalently estimate a distribution $\pi(\boldsymbol{\eta})$ for the log-usage $\boldsymbol{\eta}_{ge}$. To obtain the point estimate for $\boldsymbol{\eta}_{ge}$, we are *not* going to do the estimation individually and independently for each exon, but try to minimize the expected (with respect to the prior π) sum of squared errors, $\mathrm{E}_\pi\left[\sum_{ge}(\boldsymbol{\eta}_{ge} - \widehat{\boldsymbol{\eta}}_{ge})^2\right]$, for all exons. From the Bayes rule, we know that, given the distribution $\pi(\boldsymbol{\eta})$, the optimal $\widehat{\boldsymbol{\eta}}_{ge}$ that minimize the squared loss is the Bayes estimator (Casella and Berger, 2002):

$$\widehat{\eta}_{gei} = \frac{\int_{\mathcal{R}^2} \eta_i f\left(\boldsymbol{N}_{ge}|e^{\boldsymbol{\eta}}\right)\pi(\boldsymbol{\eta})\mathrm{d}\boldsymbol{\eta}}{\int_{\mathcal{R}^2} f\left(\boldsymbol{N}_{ge}|e^{\boldsymbol{\eta}}\right)\pi(\boldsymbol{\eta})\mathrm{d}\boldsymbol{\eta}} \text{ for } i = 1, 2. \tag{3.12}$$

On the other hand, once we know the point estimates $\widehat{\boldsymbol{\eta}}_{ge}$ for all exons, we can use smoothing

method to estimate their empirical distribution $\pi(\boldsymbol{\eta})$.

$$\widehat{\pi}(\boldsymbol{\eta}) = \frac{1}{E}\sum_{ge} K_h(\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}}_{ge}) \tag{3.13}$$

where $K_h(\cdot)$ is a 2-dimensional kernel function, the bandwidth parameter h can be chosen using a Gaussian approximation rule (Simonoff, 1996), and $E = \sum_g E_e$ is the total number of exons for all genes.

The Bayes estimator in equation (3.12) is a shrinkage estimate based on the prior $\pi(\boldsymbol{\eta})$. Of course, the prior $\pi(\boldsymbol{\eta})$ is unknown, hence the equation cannot be used directly. In practice, we can conduct the estimation by iteratively updating $\widehat{\pi}(\boldsymbol{\eta})$ in (3.13) and $\widehat{\boldsymbol{\eta}}_{ge}$ in (3.12). The detailed algorithm is given in Appendix 3.A.2. Once we obtain $\widehat{\pi}(\boldsymbol{\eta})$, we can calculate the AMAP test statistic:

$$\widehat{T}(\boldsymbol{N}_{ge}) = \frac{\int_{\exp(\boldsymbol{\eta})\in\Delta_0^{(ge)}} f\left(\boldsymbol{N}_{ge}|e^{\boldsymbol{\eta}}\right)\widehat{\pi}(\boldsymbol{\eta})\mathrm{d}\boldsymbol{\eta}}{\int_{\mathcal{R}^2} f\left(\boldsymbol{N}_{ge}|e^{\boldsymbol{\eta}}\right)\widehat{\pi}(\boldsymbol{\eta})\mathrm{d}\boldsymbol{\eta}}. \tag{3.14}$$

3.6 Simulation

We randomly selected 10,000 genes from a real RNA-seq data set (see section 3.7 for more details), and obtained a data set $\{M_{gij}, N_{geij}, L_g, L_{ge}\}$ that contains the coverages and lengths of genes and their exons as well as the total number of mappable reads S_{ij}. The dispersion parameters ϕ_{ge} for each exon were estimated by Quasi-Likelihood (QL) method (Nelder, 2000; Robinson and Smyth, 2008), and the exon usages μ_{gei} were estimated by $\frac{\sum_j N_{geij}}{\sum_j S_{ij}M_{gij}L_{ge}}$ from method of moments estimation. Then we simulated data $\tilde{N}_{geij}$ from distribution $NB(\mu_{geij}, \phi_{ge})$ where μ_{geij} were calculated using equation (3.2) and the estimated μ_{gei} for $i = 1, 2; j = 1, \cdots, J_i; g = 1, \cdots, G$ and $e = 1, \cdots, E_g$. Given the simulated exon coverages $\tilde{N}_{geij}$, we treat them as real data (assuming no knowledge of $\{\boldsymbol{\mu}_{ge}, \boldsymbol{\phi}_{ge}\}$, and test hypothesis with the proposed method in section 3.3. The noise level b_{ij} of the background was set using method in Appendix 3.A.1. This procedure was run 10 times, and the testing results were averaged across the 10 runs and reported bellow.

We first tested for expressed exons as defined in section 3.3.1. To evaluate the proposed AMAP test, we compared it with another method, which will be called *sep-test* here, proposed

by us in Wang et al. (2011). The sep-test is originally designed to test for intron retentions from one-treatment RNA-seq data, and can also be used to test for expressed exons. We used sep-test to model the exon expressions with a Gamma distribution, and under an EB framework, calculated the posterior probabilities that the exons are not expressed, i.e., below the noise level of the background for each treatment group. For any a cutoff on the posterior probabilities, we could obtain a list of expressed exons for each treatment independently, and the two lists were combined to get the positive set of expressed exons in the two treatments. Similarly, by taking the intersection of the expressed list from one treatment and the non-expressed list from another, we could also obtain a list of switch-like pattern exons as defined in section 3.3.2. The expressed or switch-like exon lists varied when changing the cutoffs. Then, the Receiver Operation Characteristic (ROC) curves the sep-test when testing for expressed exons and switch-like patterns were obtained and shown in Figure 3.2 and 3.3, respectively. By changing the controlled FDR level, we also generated ROC curves for our proposed AMAP test. We can see that the AMAP test has higher power than that of the separately testing approach.

We then tested for differential exon usages. We defined positive exons as those with fold changes exceeding a threshold FC=1.2, that is, $c = \log(1.2)$ to define the $\Delta_1^{(ge)}$ in (3.9). Another method, namely `DEXSeq` (Anders et al., 2012) was also be used to test differential usages. Since `DEXSeq` is designed to test whether $\mu_{ge1} = \mu_{ge2}$ for each exon instead of $|\log(\mu_{ge1}/\mu_{ge2})| < c$, in order to make the testing results comparable between `DEXSeq` and our AMAP test, we applied a two-stage strategy based on `DEXSeq` as introduced in section 2.3 of the Chapter 2. Specially, a list of exons was first obtained by testing equality, then the exons in the list was selected only if their estimated FC of usages exceeded the threshold of FC. In addition, `DEXSeq` could not handle exons with low counts (most of these exons had the total counts across all replicates less than 10), hence when comparing the final results, we excluded these low count exons to calculate the true/false positive rates. The ROC curves in Figure 3.4 shows that the AMAP test is more powerful than the DEXSeq method.

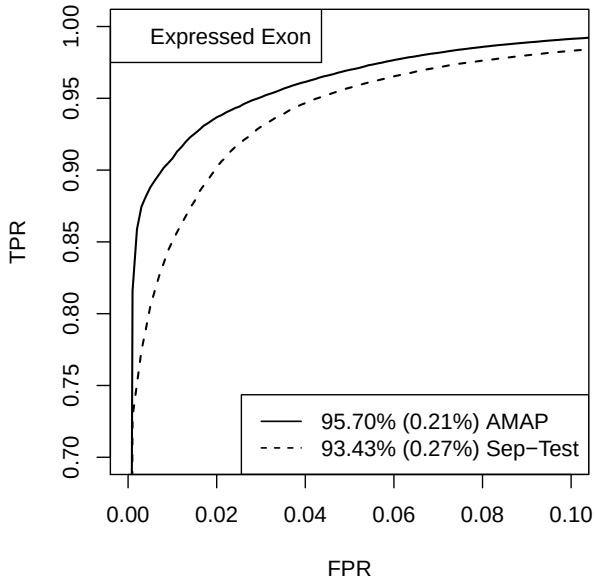


Figure 3.2: *Test for expressed exons.* On the ROC curves, each level of the TPRs were averaged across the 10 datasets. The percentage annotated for each method is the average AUC, represented as the percentage of the total area 0.1 in the range of FPR<0.1, and the percentage in each set of parentheses is the standard error of the estimated AUC in 10 runs.

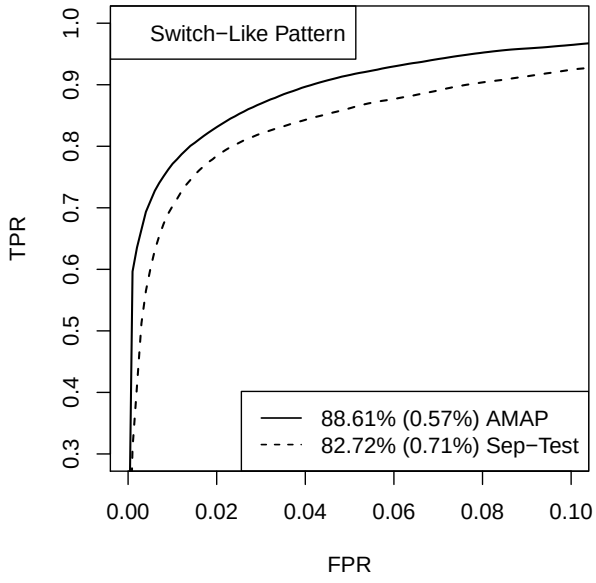


Figure 3.3: *Test for switch-like patterns.* On the ROC curves, each level of the TPRs were averaged across the 10 datasets. The percentage annotated for each method is the average AUC, represented as the percentage of the total area 0.1 in the range of FPR<0.1, and the percentage in each set of parentheses is the standard error of the estimated AUC in 10 runs.

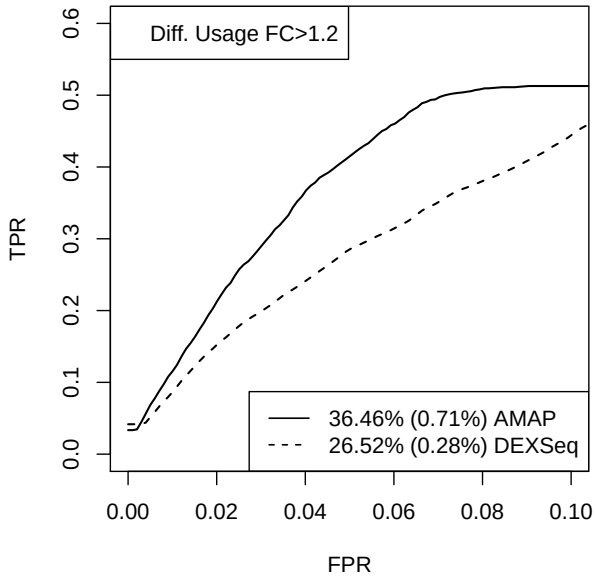


Figure 3.4: *Test for differential exon usages with FC*> 1.2. On the ROC curves, each level of the TPRs were averaged across the 10 datasets. The percentage annotated for each method is the average AUC, represented as the percentage of the total area 0.1 in the range of FPR<0.1, and the percentage in each set of parentheses is the standard error of the estimated AUC in 10 runs.

3.7 Real Data Analysis

Wang et al. (2011) investigated the gene/exon expressions at the base and tip of a rice leaf. The RNA-seq experiment included 2 treatments, from the base and tip of the leaf, with 4 replicates for each treatment. We only considered exons that are non-zero in at least one of the replicates, and the data set contained 43,110 genes and 233,538 exons in total. The noise level of background was estimated by the average read coverages on some well-defined non-coding regions (Wang et al., 2011). Then we applied several testing methods similarly to section 3.6 to test for different AS patterns, and controlled the FDR level at 1% and 5% using method in section 2.2.4 of Chapter 2 . The results are shown in Table 3.1, which shows that for each of the three patterns tested, our proposed AMAP test detected more exons than the other method in comparison.

Alternative Hypothesis	Method	FDR=1%	FDR=5%
Expressed Exons	AMAP Test	166259	184988
	Sep-Test	162205	183376
Switch-Like Patterns	AMAP Test	13174	18656
	Sep-Test	7674	11801
FC>1.2	AMAP Test	17687	64223
	DEXSeq	1211	3612

Table 3.1: *Number of positive exons from rice data.* Expressed exons, swith-like patterns and differential exon usages are detected by different method. The FDR were controlled by method in section 2.2.4 of Chapter 2 at 1% and 5%.

3.8 Discussion

We generalize the AMAP test from testing gene expression data to studying alternative splicing events from exon-level expressions. Our methods is compared with previous methods and is shown to be much more powerful. Moreover, we can easily modify the hypotheses to meet the interest of many different questions. By now, we have investigated three AS patterns such as testing for expressed exons, switch-like patterns and fold changes of exon usages. Our proposed nonparametric algorithm to estimate the distribution of exon usages is very flexible for fitting different data sets and is efficient for computation. However, noticing that when $\Delta_0^{(ge)}$ has a zero measure, for example, when testing for equality of exon usages, i.e., log-FC=0, the AMAP test statistics will always be zero, hence the algorithm needs to be modified similar to the method presented in section 2.2.5 of Chapter 2 to handle this question, and we will leave this work to future research.

3.9 APPENDICES

3.A.1 Estimation of Background

Usually, b_{ij} can be estimated by the counts and lengths of non-coding regions, or by weakly expressed genes or exons, say, the lower 5% expressed data. Suppose we have the counts N_{kij} and lengths L_{kij} for the non-coding regions indexed by $k = 1, 2, \cdots$, then we estimate

$$b_{ij} = \frac{\sum_k N_{kij}}{S_{ij} \sum_k L_{kij}}$$

3.A.2 Estimation of Prior Distribution

The following algorithm can be employed to estimate the prior distribution $\pi(\boldsymbol{\eta})$ for $\boldsymbol{\eta}_{ge} = (\eta_{ge1}, \eta_{ge2})$ in section 3.5.

Algorithm 3.1. *Estimate the prior distribution*

1. *Get an initial estimate of* η_{gei}*, for example, by* $\widehat{\eta}_{gei} = \log(\widehat{\mu}_{gei})$ *for* $\widehat{\mu}_{gei} = \texttt{mean}_j \left\{ \frac{N_{geij}}{S_{ij} M_{gij}} \right\}$ *from equation (3.2) by method of moment estimation (MME).*

2. *Approximate* $\pi(\boldsymbol{\eta})$ *using kernel smoothing method by equation 3.13. We can use 2-dimensional Gaussian kernel* $K_h(\cdot)$*, and the bandwidth parameter* h *can be chosen using a Gaussian approximation rule (Simonoff, 1996).*

3. *Use the estimate* $\widehat{\pi}(\boldsymbol{\eta})$ *from the previous step as the prior, and calculate the Bayes estimator for each* $\boldsymbol{\eta}_{ge}$ *by equation (3.12)*

4. *Repeat step 2-3 until the estimate* $\widehat{\pi}(\boldsymbol{\eta})$ *becomes stable.*

In the last step of the algorithm, we need to calculated the difference between the two estimates $\pi(\boldsymbol{\eta})$ before and after the update. Denote the estimates from two subsequent iterations by $\widehat{\pi}$ and $\widehat{\pi}^*$, then we can calculate the Cramér-von Mises criterion (Anderson, 1962) $D = \int_{\mathcal{R}^2} [F(\boldsymbol{\eta}) - F^*(\boldsymbol{\eta})]^2 \widehat{\pi}(\boldsymbol{\eta}) \mathrm{d}\boldsymbol{\eta}$, where F and F^* are the cumulative distribution function (CDF) of $\widehat{\pi}$ and $\widehat{\pi}^*$, respectively, and a small value of D indicates insignificantly changes between the two estimated distributions, hence we can stop the iteration. By our experience, this algorithm usually converges in less than five iterations.

CHAPTER 4. Model-Based Clustering for RNA-seq Data

Yaqing Si and Peng Liu

Iowa State University, Snedecor Hall, Ames, IA 50011, USA.

Pinghua Li, and Thomas Brutnell

Brutnell Lab, Donald Danforth Plant Science Center, St. Louis, MO 63132, USA.

This work was submitted to *Annals of Applied Statistics*

Abstract

RNA-seq technology has been widely adopted as an attractive alternative to microarray-based methods to study global gene expression. However, robust statistical tools to analyze these complex datasets are lacking. By grouping genes with similar expression profiles across treatments, cluster analysis provides insight into gene functions and networks and hence is an important technique for RNA-seq data analysis. In this manuscript, we derive clustering algorithms based on appropriate probability models for RNA-seq data. An Expectation-Maximization (EM) algorithm and another two stochastic methods are described. In addition, a strategy for initialization based on likelihood is proposed to improve the clustering algorithms. Moreover, we present a model-based hybrid-hierarchical clustering method to generate a tree structure that allows visualization of relationships among clusters as well as flexibility of choosing the number of clusters. Results from both simulation studies and analysis of a maize RNA-seq data set show that our proposed methods provide better clustering results than alternative methods such as the K-means algorithm and hierarchical clustering methods that are not based on probability models.

Key Words: Model-based clustering; RNA-seq; Expectation-Maximization (EM) algorithm; simulated annealing; deterministic annealing.

4.1 Introduction

Next-generation sequencing (NGS) technologies have been revolutionizing studies of genome structure, gene expression and epigenetics (Metzker, 2010; Wang, Li and Brutnell, 2010). One important application of NGS technologies is in the study of gene expression by measuring messenger RNA (mRNA) levels for all genes in a sample. This technology is called RNA-seq, and several recent reviews have described this nascent technology (Metzker, 2010; Wang, Li and Brutnell, 2010; Wang, Gerstein and Snyder, 2009; Marguerat, Wilhelm and Bähler, 2008). Here we briefly describe how RNA-seq data can be generated. The complete set of mRNA molecules are first extracted from a sample and converted to a library of short complementary DNA (cDNA) fragments. Then these fragments are sequenced simultaneously by NGS technology. The resulting millions of short sequences, which are commonly called reads, are then aligned to a reference genome or reference transcripts. Gene expression is measured by the enumeration of reads mapped to each gene where the gene can be defined as a collection of exons or other appropriate definitions given context of a study (Bullard et al., 2010). The resulting RNA-seq data are essentially digital signals that can be used to quantify levels of gene expression (Marguerat, Wilhelm and Bähler, 2008; Wang, Gerstein and Snyder, 2009). This differs from microarray technologies which measure gene expression by fluorescence intensities detected from hybridized samples. Inescapable factors such as cross-hybridization, secondary structure of the DNA and technical challenges associated with fluorescent detection used in microarray analysis limit both the sensitivity and dynamic range. Compared with microarray technologies, NGS technologies permit quantitative measures of gene expression over a much larger dynamic. These advantages have rapidly accelerated the adoption of the NGS technologies in studies of gene expression and present new challenges to data analysis.

In the pioneering studies using RNA-seq, only two treatment groups were analyzed Sultan et al. (2008); Marioni et al. (2008). More recently, RNA-seq experiments that examined multiple treatment groups have been published. For example, Li et al. (2010) carefully selected a developing leaf from a corn plant that captures multiple stages of photosynthetic differentiation. They exploited Illumina sequencing technologies to profile gene expression from four represen-

tative sections of the leaf blade. One major goal of this study was to survey gene expression profiles along different developmental stages to gain understanding of the transcriptional network associated with the development of C4 photosynthesis. In this endeavor, cluster analysis is an important tool as it often reveals groups of genes with similar expression patterns, where genes within such groups tend to be functionally related.

Li et al. (2010) took an heuristic approach by applying the K-means algorithm to partition log-transformed data for the differentially expressed genes. The K-means algorithm starts from an initial partition of the objects (genes) and proceeds by iteratively calculating the centers (means) of clusters and reassigning each object to the closest cluster according to some measurement of distance such as Euclidean distance. This iteration continues until no more reassignments take place. Although this heuristic approach is easy to implement, its performance was not evaluated for RNA-seq data analysis. Studies of clustering algorithms with microarray data revealed that heuristic algorithms performed worse than model-based algorithms (Yeung et al, 2001). Surprisingly, there has been no published statistical research to examine cluster analysis of RNA-seq data although it is urgently needed due to the huge amount of data being generated. In this paper, we address this need by deriving model-based clustering algorithms based on appropriate probability models for RNA-seq data and evaluating the performance of the model-based approach and heuristic algorithms including the K-means method.

RNA-seq data have been modeled using a Poisson (Bullard et al., 2010; Marioni et al., 2008) or negative binomial distribution (Robinson and Oshlack, 2010). We describe the two distributions in section 4.2 and show how our model-based clustering method handles both probability models in a unified fashion. We present an Expectation-Maximization algorithm for estimating the model parameters and cluster membership in section 4.3.1. In addition, a model-based initialization algorithm is proposed in section 4.3.2 to reduce the dependence on the initialization. We also describe two stochastic versions of EM algorithms in section 4.3.3 that are intended to reduce the chance of being trapped at local solutions. A model-based hierarchical algorithm is proposed in section 4.3.4 to generate a hierarchical structure of the clusters and allow more flexibility of choosing cluster numbers. In section 4.4, we simulate data and compare

the proposed method with others using three commonly used criteria: sensitivity, specificity and mutual information (Booth, Casella and Hobert, 2008; Woodard and Goldszmidt, 2011; Strehl and Ghosh, 2002). In section 4.5, we apply the model-based method to the data from Li et al. (2010), and evaluate our results by comparing the clusters with gene annotations. We state our conclusion in section 4.6 that our results from extensive simulation studies and an analysis of an RNA-seq dataset all show that our proposed method outperforms alternative methods, namely, the K-means algorithm and self-organizing map (SOM) (Tamayo et al, 1999; Ressom and Natarajan, 2003).

4.2 Model

Let N_{gij} denote the count of reads mapped to gene g for replicate j of treatment i for $g = 1, \cdots, G; i = 1, \cdots, I;\ j = 1, \cdots, n_i$, where G is the total number of genes of interest, I is the number of treatment groups, and n_i is the number of replicates for treatment i. Two discrete probability distributions have been proposed to model RNA-seq data. The Poisson distribution has been shown to be appropriate for the RNA-seq data when technical replicates are performed (Marioni et al., 2008; Bullard et al., 2010). When there are biological replicates, RNA-seq data may exhibit more variability than expected with a Poisson distribution, i.e., the overdispersion phenomenon (Anders and Huber, 2010). The negative binomial (NB) model proposed by Robinson and Smyth (2008) originally for serial analysis of gene expression (SAGE) data allows overdispersion and has been applied to RNA-seq data analysis (Robinson and Oshlack, 2010; Anders and Huber, 2010). We consider both distributions in this paper.

4.2.1 Poisson Distribution

Suppose N_{gij} follows a Poisson distribution with mean λ_{gij} that is parameterized as:

$$\log \lambda_{gij} = s_{gij} + \alpha_g + \beta_{gi} \tag{4.1}$$

with $\sum_{i=1}^{I} \beta_{gi} = 0$. The offset term s_{gij} is a normalization factor that may depend on the gene length and library of a sample such as the total number of mapped reads of a library. Once estimated from data, the normalization factor is often treated as known in the model (Bullard

et al., 2010; Robinson and Oshlack, 2010; Marioni et al., 2008). The parameter α_g represents the mean expression level of gene g across all treatments; β_{gi} measures the expression level of gene g in treatment i relative to the overall mean expression. To cluster gene expression profiles, we are interested in clustering the vectors $\boldsymbol{\beta}_g = (\beta_{g1}, \cdots, \beta_{gI})$ for all G genes.

4.2.2 Negative Binomial Distribution

For the negative binomial (NB) model, we adopt the parameterization in Robinson and Smyth (2008) by modeling the variance as

$$\text{Var}(N_{gij}) = \lambda_{gij} + \phi_g \lambda_{gij}^2, \tag{4.2}$$

where λ_{gij} is the same as in (4.1) and ϕ_g is a dispersion parameter. Compared with Poisson model, an extra parameter, ϕ_g, is introduced for each gene. Robinson and Smyth (2008) described several methods to estimate ϕ_g. In this paper, we estimate ϕ_g by the quasi-likelihood (QL) method. To simplify the algorithm, we treat ϕ_g as known upon its estimation because our numerical studies showed this strategy produced similar clustering results to those based on the true ϕ_g values (see section 4.4.3). With this strategy, the unknown parameters are the same for the Poisson and NB models and thus we denote the likelihood function for both models by $f(\boldsymbol{N}_g|\alpha_g, \boldsymbol{\beta}_g)$ for gene g where $\boldsymbol{N}_g = \{N_{gij}\}$.

4.3 Model-Based Clustering

Model-based clustering methods assume that data are generated by a mixture of probability distributions where each component corresponds to one cluster. Extensive research has been done in model-based clustering with multivariate normal mixture distributions. See, for example, Fraley and Raftery (2002) for an excellent review. In this section, we describe model-based clustering for RNA-seq data with the probability models introduced in section 4.2.

The algorithms described below aim to cluster gene expression profiles, which is desired in practical application. Consequently, genes within the same cluster have similar expression profiles (denoted by $\boldsymbol{\beta}_g$ in our notation), but may have different overall mean expression levels

(indicated by α_g). However, it is straightforward to make changes in the algorithm if the goal is to cluster according to both the overall expression levels and the expression profiles $\alpha_g + \boldsymbol{\beta}_g$.

Suppose there are K clusters and let $\boldsymbol{\mu}_k = (\mu_{k1}, \cdots, \mu_{kI})$ denote the *center* of cluster k with $\sum_{i=1}^{I} \mu_{ki} = 0$ for $k = 1, \cdots, K$. The likelihood of the mixture model for gene g is $\sum_k p_k f(\boldsymbol{N}_g | \alpha_g, \boldsymbol{\beta}_g = \boldsymbol{\mu}_k)$, where $f(\boldsymbol{N}_g | \alpha_g, \boldsymbol{\beta}_g = \boldsymbol{\mu}_k)$ is the likelihood if gene g belongs to the kth cluster and p_k is the mixing proportion with $p_k \geq 0$ and $\sum_{k=1}^{K} p_k = 1$. The likelihood function can be based on a Poisson model or NB model as described in section 4.2. Taking all genes together, the likelihood is:

$$\prod_g \sum_k p_k f(\boldsymbol{N}_g | \alpha_g, \boldsymbol{\beta}_g = \boldsymbol{\mu}_k).$$

Note that we assume independence among genes which is likely not true in real situations. However, it is difficult, or impossible, to model and estimate the correlation among tens of thousands of genes with only several replicates and no prior knowledge about the relationship among genes. Thus, for simplicity, we take the independence assumption for simplicity as in previous model-based cluster analysis for microarray studies (Yeung et al, 2001).

4.3.1 Model-Based Clustering with the Expectation-Maximization Algorithm (MB-EM)

The EM algorithm has been widely applied to model-based clustering with multivariate normal mixture distributions (Fraley and Raftery, 2002). Similarly, we derive an EM algorithm (Algorithm 4.1) for RNA-seq data with a mixture of Poisson or NB models. Let $Z_{gk} = 1$ if gene g belongs to the kth cluster and $Z_{gk} = 0$ otherwise. The EM algorithm views the cluster memberships $\boldsymbol{Z} = \{Z_{gk} : g = 1, \cdots, G; k = 1, \cdots, K\}$ as missing data and proceeds by iteratively calculating the conditional expectations of $\boldsymbol{Z}$ and updating the estimates for model parameters until convergence:

Algorithm 4.1 (MB-EM Algorithm)**.**

(i) Initialization*: Set* $p_k^{(1)}$ *according to prior knowledge about the cluster size. If no such information is available, let* $p_k^{(1)} = 1/K$ *for* $k = 1, \cdots, K$*. Choose* K *vectors* $\boldsymbol{\mu}_1^{(1)}, \cdots, \boldsymbol{\mu}_K^{(1)}$

with $\sum_{i=1}^{I} \mu_{ki}^{(1)} = 0$ *for* $k = 1, \cdots, K$ *as the initial set of cluster centers. See Algorithm 4.2 for one way to choose these* $\boldsymbol{\mu}_k^{(1)}$. *Obtain the initial values of* $\boldsymbol{\alpha}^{(1)} = \{\alpha_{gk}^{(1)} : g = 1, \cdots, G; k = 1, \cdots, K\}$ *by maximizing* $f(\boldsymbol{N}_g | \alpha_{gk}, \boldsymbol{\mu}_k^{(1)})$ *with respect to* α_{gk} *for each combination of gene* g *and cluster* k.

(ii) E-step*: Calculate the conditional expectation of* Z_{gk} *given data and parameters estimated from the* m*th step* $(\boldsymbol{\mu}^{(m)}, \boldsymbol{p}^{(m)}, \boldsymbol{\alpha}^{(m)})$, *where* $\boldsymbol{\mu}^{(m)} = \{\boldsymbol{\mu}_k^{(m)} : k = 1, \cdots, K\}, \boldsymbol{p}^{(m)} = \{p_k^{(m)} : k = 1, \cdots, K\}, \boldsymbol{\alpha}^{(m)} = \{\alpha_{gk}^{(m)} : g = 1, \cdots, G; k = 1, \cdots, K\}$. *To simplify notation, we use* $\widehat{Z}_{gk}$ *to denote the conditional expectation:*

$$\widehat{Z}_{gk}^{(m)} = \mathrm{E}(Z_{gk} | \boldsymbol{N}, \boldsymbol{\mu}^{(m)}, \boldsymbol{p}^{(m)}, \boldsymbol{\alpha}^{(m)}) = \frac{p_k^{(m)} f(\boldsymbol{N}_g | \alpha_{gk}^{(m)}, \boldsymbol{\mu}_k^{(m)})}{\sum_l p_l^{(m)} f(\boldsymbol{N}_g | \alpha_{gl}^{(m)}, \boldsymbol{\mu}_l^{(m)})}. \tag{4.3}$$

(iii) M-step*: Update the parameter estimates by*

$$\boldsymbol{\mu}_k^{(m+1)} = \underset{\{\sum_i \mu_{ki} = 0\}}{\operatorname{argmax}} \sum_g \widehat{Z}_{gk}^{(m)} \log f\left(\boldsymbol{N}_g \middle| \alpha_{gk}^{(m)}, \boldsymbol{\mu}_k\right),$$

$$p_k^{(m+1)} = \sum_g \widehat{Z}_{gk}^{(m)} / G,$$

and

$$\alpha_{gk}^{(m+1)} = \underset{\alpha_{gk}}{\operatorname{argmax}} f\left(\boldsymbol{N}_g | \alpha_{gk}, \boldsymbol{\mu}_k^{(m+1)}\right),$$

where $\widehat{Z}_{gk}^{(m)}$ *is obtained from from step (ii).*

(iv) Return to step (ii) or stop the iteration if change of the total log-likelihood is small.

(v) For each $g = 1, \cdots, G$, *assign gene* g *to cluster* k *if* $k = \operatorname{argmax}_l \widehat{Z}_{gl}$, *where* $\widehat{Z}_{gl}$ *is obtained after the convergence of above steps.*

Note that Algorithm 4.1 not only assigns gene g to cluster k but also provides a measure of the uncertainty in the assignment by $1 - \widehat{Z}_{gk}$. If clustering based on $\alpha_g + \boldsymbol{\beta}_g$ is preferred, then we don't estimate α_{gk} but estimate α_k together with μ_k and corresponding calculations in step (i)-(iii) can be easily modified.

4.3.2 Initialization

It is well known that initialization of the cluster centers impacts both the speed of convergence and the outputs of the EM algorithm (Park, Yoo and Cho, 2005; Hall, Özyur and Bezdek, 1999; Fraley and Raftery, 2002). To tackle this problem, Arthur and Vassilvitskii (2007) proposed to pick the initial cluster centers from observations in a specific way such that they are well separated from each other with respect to some distance measure. Following this idea, rather than choosing K genes uniformly at random from all genes and using their expression profiles as the initial cluster centers, we only choose one cluster center uniformly at random and then set the additional centers gradually by selecting genes based on the distance between each gene and each of the selected centers. Here, the distance is measured by likelihood function.

Algorithm 4.2 (Model-based Initialization for Cluster Centers)**.**

(i) Choose one gene randomly from all genes, and set the initial center for cluster 1, $\boldsymbol{\mu}_1^{(1)}$*, to be the maximum likelihood estimate (MLE) of* $\boldsymbol{\beta}_g$ *of the selected gene.*

(ii) Given m *center(s),* $\boldsymbol{\mu}_1^{(1)}, \cdots, \boldsymbol{\mu}_m^{(1)}$ *for* $1 \leq m < K$*, selected from previous steps, calculate the measure of the distance,* d_{gl}*, between each gene* g *and each previously selected cluster center* $\boldsymbol{\mu}_l^{(1)}$ *by*

$$d_{gl} = \log \frac{\max_{\alpha_g \in \mathcal{R}, \sum \beta_{gi}=0} f(\boldsymbol{N}_g | \alpha_g, \boldsymbol{\beta}_g)}{\max_{\alpha_g \in \mathcal{R}} f(\boldsymbol{N}_g | \alpha_g, \boldsymbol{\beta}_g = \boldsymbol{\mu}_l^{(1)})},$$

for $g = 1, \cdots, G; l = 1, \cdots, m$*. Then randomly select a gene with probability* $p_g = d_g^2 / \sum_{g'=1}^{G} d_{g'}^2$ *for* $d_g = \min\{d_{g1}, \cdots, d_{gm}\}$ *and set a new center* $\boldsymbol{\mu}_{m+1}^{(1)}$ *as the MLE of* $\boldsymbol{\beta}_g$ *for the selected gene in this step.*

(iii) Repeat step (ii) until K *cluster centers are obtained.*

By the definitions of d_g and p_g in step (ii) of Algorithm 4.2, a gene is more likely to be selected if it is far away from all existing centers. Hence the K centers chosen by this algorithm are expected to be separated better than a set of centers that are randomly selected. Our simulation study shows that this algorithm improves the performance of EM algorithm (section 4.4.4).

4.3.3 Other Algorithms for Model-Based Clustering

The EM algorithm does not guarantee global optimal solutions. Several stochastic algorithms have been proposed to reduce the the risk of being trapped in local solutions. We describe two in this subsection and will examine their performances in our analysis. Both algorithms modify formula (4.3) to calculating $\widehat{Z}_{gk}^{(m)}$ in step (ii) of Algorithm 4.1.

(a) According to the deterministic annealing (DA) algorithm described in Rose (1998), the cluster in the mth iteration step is updated by

$$\widehat{Z}_{gk}^{(m)} = \frac{p_k^{(m)}\{f(\boldsymbol{N}_g|\alpha_{gk}^{(m)}, \boldsymbol{\mu}_k^{(m)})\}^{1/\tau_m}}{\sum_l p_l^{(m)}\{f(\boldsymbol{N}_g|\alpha_{gl}^{(m)}, \boldsymbol{\mu}_l^{(m)})\}^{1/\tau_m}}. \tag{4.4}$$

(b) The classification EM (CEM) algorithm with simulated annealing (SA) proposed by Celeux and Govaert (1992) updates the estimate of Z_{gk} by

$$\widehat{Z}_{gk}^{(m)} = \frac{\{p_k^{(m)} f(\boldsymbol{N}_g|\alpha_{gk}^{(m)}, \boldsymbol{\mu}_k^{(m)})\}^{1/\tau_m}}{\sum_l \{p_l^{(m)} f(\boldsymbol{N}_g|\alpha_{gl}^{(m)}, \boldsymbol{\mu}_l^{(m)})\}^{1/\tau_m}}. \tag{4.5}$$

Both algorithms employ the annealing procedure with a sequence of preselected annealing rates ("temperatures", τ_m) decreasing to zero from a positive number. Apparently, when fixing $\tau_m = 1$, both algorithm updates the values of $\widehat{Z}_{gk}^{(m)}$ the same way as the EM algorithm. Hence, the Algorithm 4.1 can be viewed as a special case with a constant annealing rate $\tau_m \equiv 1$. As $\tau_m \to \infty$, we always get $\widehat{Z}_{gk}^{(m)} = p_k$ for DA algorithm and $1/K$ for SA algorithm, which means that genes are assigned to each cluster totally randomly. On the other hand, as $\tau_m \to 0$ the randomness is gradually lost and we finally get $Z_{gk} = 0$ or 1, i.e, a hard cluster solution. Hence, τ_m determines the amount of randomness added in each step while searching for solutions. To apply these algorithms, we follow the suggestions of Rose (1998) and use $\tau_{m+1} = 0.9\tau_m$ with $\tau_1 = 2$.

For the SA algorithm proposed in Celeux and Govaert (1992), another difference from the EM algorithm (Algorithm 4.1) is that, before updating parameter values in the M-step, each gene is assigned to a cluster based on one observation simulated from a multinomial distribution with probabilities $\widehat{Z}_{gk}^{(m)}$ as calculated by equation (4.5).

4.3.4 Model-Based Hybrid-Hierarchical Clustering Algorithm (MB-HH)

So far, we have assumed that the number of clusters, K, is predetermined. For a real data analysis, this quantity often needs to be estimated. There are different methods that can be applied to estimating K. For instance, choose the K that minimizes the Bayesian information criterion (BIC) for the mixture model. Alternatively, instead of choosing a single value of K for the clustering analysis, we can build a hierarchical tree of clusters. The hierarchical structure of the clusters provides information about the relationships of clusters and allows flexibility of obtaining different number of clusters by cutting the tree at different levels.

There can be tens of thousands of genes from RNA-seq data to cluster, and treating each gene as the smallest cluster at the bottom of the tree requires intensive computation. To speed up the calculation, we propose to use agglomerative (bottom-up) strategy starting with K_0 clusters, where K_0 is a number relatively large to allow enough resolution but far less than the number of genes, G. The initial K_0 clusters can be obtained by the model-based clustering algorithms described in the previous subsections. In each of the following steps, two clusters are merged if the 'distance' between them is the smallest among all possible pairs. Finally after $K_0 - 1$ steps, all genes belong to a single cluster and the hierarchical tree is built up. Such an algorithm has been called hybrid-hierarchical (HH) clustering algorithm (Vaithyanathan and Dom, 2000; Zhong and Ghosh, 2003). Here, the term 'hybrid' is used to point out that the HH algorithm combines the starting steps that obtain K_0 clusters using non-hierarchical methods and the merging steps that are similar to ordinary hierarchical clustering.

After the mth $(0 \leq m < K_0)$ merging step, we denote the $K_0 - m$ clusters by disjoint sets $\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_{K_0-m}$, and calculate the distance between two clusters, say $\mathcal{G}_k$ and $\mathcal{G}_l$, by the following formula :

$$D(\mathcal{G}_k, \mathcal{G}_l) = \sum_{g \in \mathcal{G}_k} \log \frac{f(\boldsymbol{N}_g | \alpha_g^{(k)}, \boldsymbol{\mu}_k)}{f(\boldsymbol{N}_g | \alpha_g^{(kl)}, \boldsymbol{\mu}_{(kl)})} + \sum_{g \in \mathcal{G}_l} \log \frac{f(\boldsymbol{N}_g | \alpha_g^{(l)}, \boldsymbol{\mu}_l)}{f(\boldsymbol{N}_g | \alpha_g^{(kl)}, \boldsymbol{\mu}_{(kl)})}, \tag{4.6}$$

where $\alpha_g^{(k)}$ and $\boldsymbol{\mu}_k$ maximize the likelihood $f(\boldsymbol{N}_g | \alpha_g, \boldsymbol{\mu}_k)$, and $\boldsymbol{\mu}_{(kl)}$ is the center of the cluster formed by merging $\mathcal{G}_k$ and $\mathcal{G}_l$. This distance is the reduction of total log-likelihood from before to after the mergence. Obviously, merging clusters with the minimal distance defined in (4.6)

aims to achieve the maximum log-likelihood in each step (Fraley, 1999; Meila and Heckerman, 2001).

4.4 Simulation Study

We conducted simulation studies to compare model-based clustering methods with other methods, including K-means and SOM, which have been popularly used in microarray data analysis and could also be applied to analyzing RNA-seq data. We first describe the way data was generated in section 4.4.1 and present the criteria used to evaluate the clustering performance in section 4.4.2. Then we check the validity of treating the estimated dispersion parameter ϕ_g as known for NB models in section 4.4.3 and evaluate the model-based initialization algorithm (Algorithm 4.2) versus random initialization in section 4.4.4. Finally, in section 4.4.5, we compare our proposed algorithms with others.

4.4.1 Data simulation

We considered an experiment with three treatment groups and three replicates for each treatment group. This is a case easily encountered in real data analysis. Suppose that there were $K = 7$ different expression patterns across three treatments and the cluster centers were characterized by $\boldsymbol{\mu}_k = \eta_\mu \boldsymbol{\delta}_k$, where η_μ determined the magnitude of gene expression changes across treatments and $\boldsymbol{\delta}_k = (\delta_{k1}, \delta_{k2}, \delta_{k3})$ described the pattern of changes for cluster k, for $k = 1, \cdots, K$. A larger η_μ means larger distances between the centers and better separation of clusters. The distinct profiles characterized by $(\delta_{k1}, \delta_{k2}, \delta_{k3})$ are listed bellow:

cluster k	1	2	3	4	5	6	7
δ_{k1}	-1	-1	0	0	1	1	0
δ_{k2}	0	1	-1	1	-1	0	0
δ_{k3}	1	0	1	-1	0	-1	0

.

For the first cluster, the expression of genes increases from the first treatment group to the second one and increases further for the third treatment group. For the second cluster, the expression increases from first treatment group to the second one but then decreases for the

third group. Note that the last cluster has a mean profile identically zero and this cluster corresponds to the group of genes that are non-differentially expressed (non-DE) across treatments. Although only identified differentially expressed (DE) genes are typically included in the cluster analysis, there could be false positives on the list of identified genes. For the simulation study, we included this cluster of non-DE genes to make our simulation more general and did not expect this to affect the relative ranking of the evaluated methods.

RNA-seq data for $G = 1000$ genes were simulated for each dataset according to the following regime. For each $g = 1, \cdots, G$, $\boldsymbol{Z}_g^0 = \{Z_{gk}^0 : k = 1, \cdots, 7\}$ was drawn independently from a multinomial distribution with equal probabilities, where $Z_{gk}^0 = 1$ means gene g belongs to cluster k and $Z_{gk}^0 = 0$ otherwise. Given $Z_{gk}^0 = 1$, the gene expression profile was simulated according to $\boldsymbol{\beta}_g = \boldsymbol{\mu}_k + \boldsymbol{\epsilon}_g$, where $\boldsymbol{\mu}_k = \eta_\mu \boldsymbol{\delta}_k$ as described above and $\boldsymbol{\epsilon}_g = (\epsilon_{g1}, \epsilon_{g2}, \epsilon_{g3})$ added fluctuation around cluster center $\boldsymbol{\mu}_k$ specifically for gene g. We sampled ϵ_{gi} for $i = 1, 2, 3$ from $\eta_\mu \eta_\epsilon N(0, 0.2^2)$, where η_ϵ controlled the level of fluctuation relative to the cluter center $\eta_\mu \boldsymbol{\delta}_k$. The overall mean expression level α_g was drawn from $\eta_\alpha N(4, 1)$, where η_α controlled the magnitude of average expression level. The dispersion parameter ϕ_g was simulated from $\eta_\phi \text{Gamma}(0.75, 2)$, where $\text{Gamma}(0.75, 2)$ is a gamma distribution with mean $0.75/2$ and variance $0.75/2^2$. Changing the value of η_ϕ allowed different levels of dispersion. Specially, $\eta_\phi = 0$ corresponds to the Poisson model, which is the limiting case of NB model as the dispersion approaches zero. The normalization factor s_{gij} was generated from $N(0, 1)$. Given these parameters, the gene expression count N_{gij} was generated from the NB model with expectation $\exp(s_{gij} + \alpha_g + \beta_{gi})$ and dispersion ϕ_g.

Once the data set was simulated, we treated all parameters except s_{gij} as unknown to resemble a real experiment. The values of $\eta_\mu, \eta_\epsilon, \eta_\alpha$ and η_ϕ were varied to create different simulation settings, and 50 data sets were independently simulated for each setting.

4.4.2 Assessment of performance

We assessed the performances of different clustering approaches by comparing the resulting partitions with the original partition of genes defined by $\boldsymbol{Z}^0 = \{\boldsymbol{Z}_g^0 : g = 1, \cdots, 1000\}$. A better performance is indicated by more agreement between the two partitions. The following

three statistics were used to evaluate the agreement. For all the three statistics, higher values indicate better performance.

1. *Pairwise Sensitivity*: the proportion of pairs of genes (objects) that are clustered together among all pairs that had the same original assignment (Booth, Casella and Hobert, 2008; Woodard and Goldszmidt, 2011).

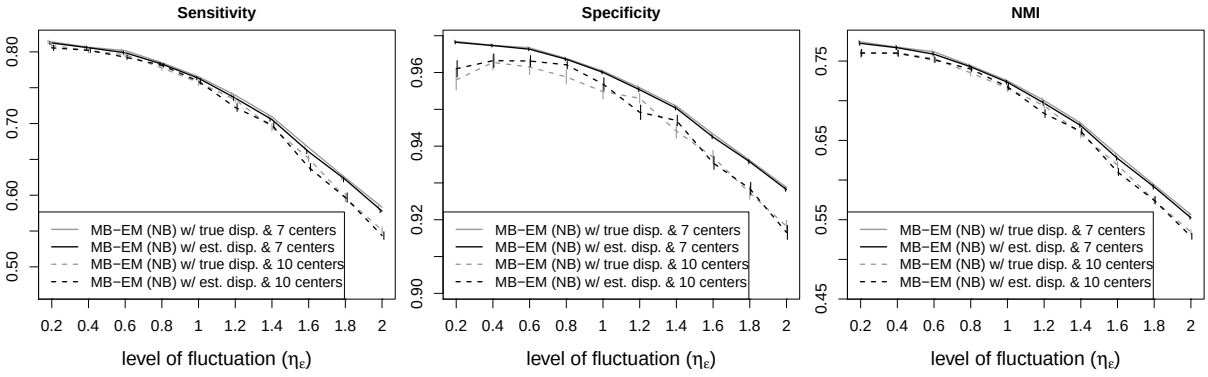
2. *Pairwise Specificity*: the proportion of pairs of genes (objects) that are clustered to different groups among all pairs that had different original assignment (Booth, Casella and Hobert, 2008; Woodard and Goldszmidt, 2011).

3. *Normalized Mutual Information (NMI)*: Mutual information (MI) is used in information theory to measure the amount of information one random variable contains about another, or equivalently, the reduction in the uncertainty of one due to the knowledge of the other. Here, MI is used to quantify the shared information between the true partition and the clustering result. See Strehl and Ghosh (2002) for the explicit formula for calculation using the contingency table formed by the two partitions. MI value is high if there is strong dependence (more shared information) between the two partitions, and is close to zero otherwise. Since there is no upper bound for MI, its normalized version ranging from 0 to 1 is often desirable for easier comparison (Strehl and Ghosh, 2002).

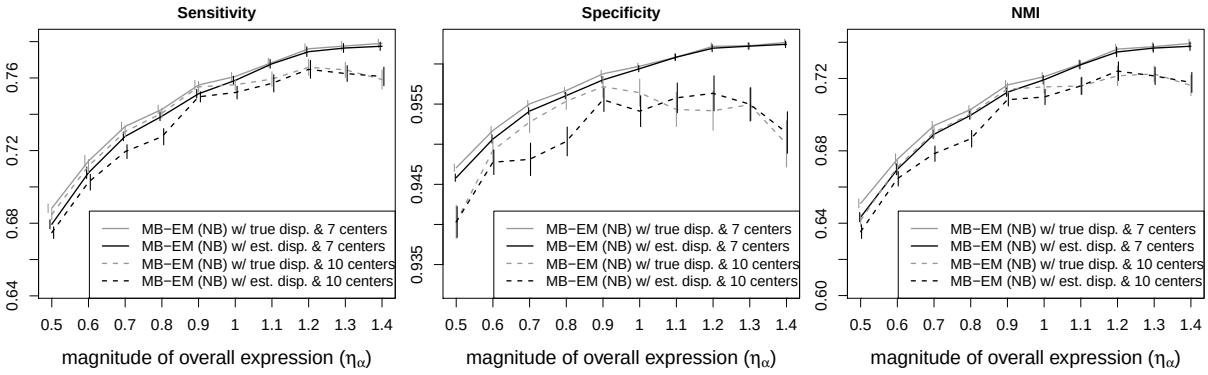
4.4.3 Validation of Estimating Dispersion Parameters

We estimated the dispersion parameters ϕ_g and treated them as if they were true values when applying the model-based clustering algorithms. However, it is challenging to obtain good estimates of dispersion parameters due to the small number of replicates in RNA-seq data. To examine the impact of the estimated parameters on cluster analysis, we compared the model-based clustering methods using estimated values for ϕ_g versus that using the input (true) values employed to simulate the counts.

Figure 4.1(a) plots the values of sensitivity, specificity, and NMI for different clustering approaches over a range of η_ϵ values used to simulate RNA-seq data while other parameters η_μ, η_α and η_ϕ were fixed at 1. All three statistics decrease as the level of gene-specific fluctuation



(a)



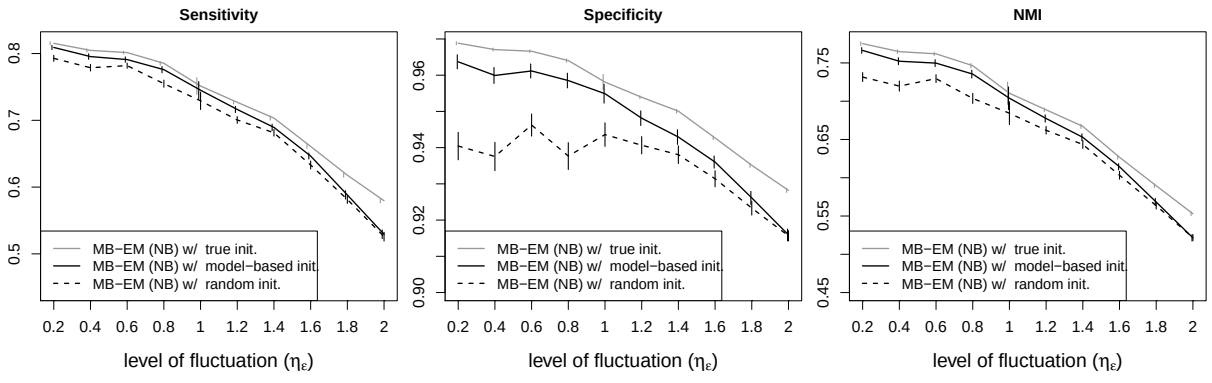
(b)

Figure 4.1: *Estimation of dispersion parameters.* Clustering results from the MB-EM algorithms using true and estimated dispersion parameters were compared. For each parameter setting, result from 50 data sets were averaged and plotted on the line. The length of vertical bars represents standard error. Solid lines are for results of 7 clusters and dashed lines are for results of 10 clusters.

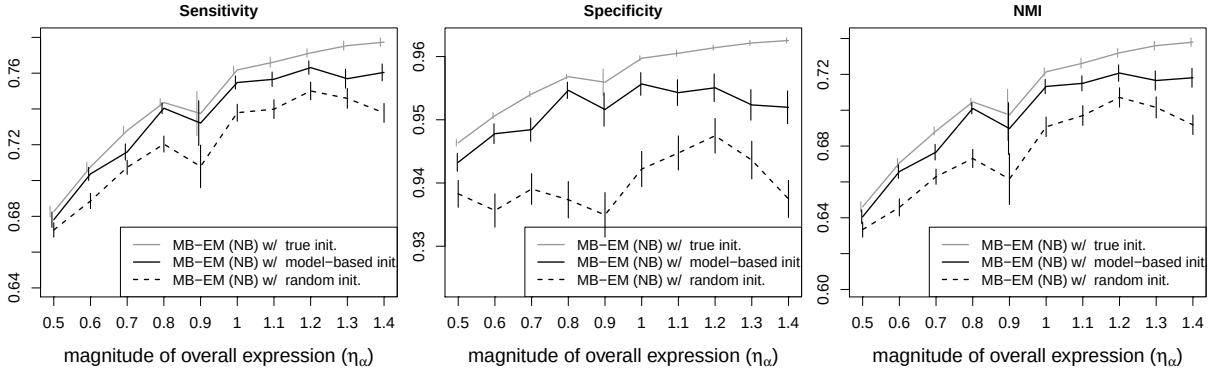
around cluster centers, η_ϵ, increases. Solid lines correspond to results with $K = 7$, the true number of clusters used to simulate data. The MB-EM algorithms using true and estimated dispersions perform indistinguishably as shown in Figure 4.1(a). In practice, the true number of clusters is unknown and we might apply a different number in cluster analysis. Hence, we also did cluster analysis with $K = 10$. Still, the clustering results from using true and estimated dispersions are almost the same (see Figure 4.1(a)). We also varied parameters η_α, η_μ and η_ϕ one at a time while keeping others fixed at 1 to generate RNA-seq datasets. The results are shown in Figure 4.1(b) and 4.A.1. The biggest difference between using true and estimated dispersions was observed at parameter setting $\eta_\alpha = 0.8$ and $\eta_\epsilon = \eta_\mu = \eta_\phi = 1$ (see Figure 4.1(b)), while

the differences were much smaller at most of the other parameter settings. Consequently, all results presented later were obtained using estimated dispersion parameters just like how we analyze real data.

4.4.4 Comparison of Initialization Algorithms



(a)



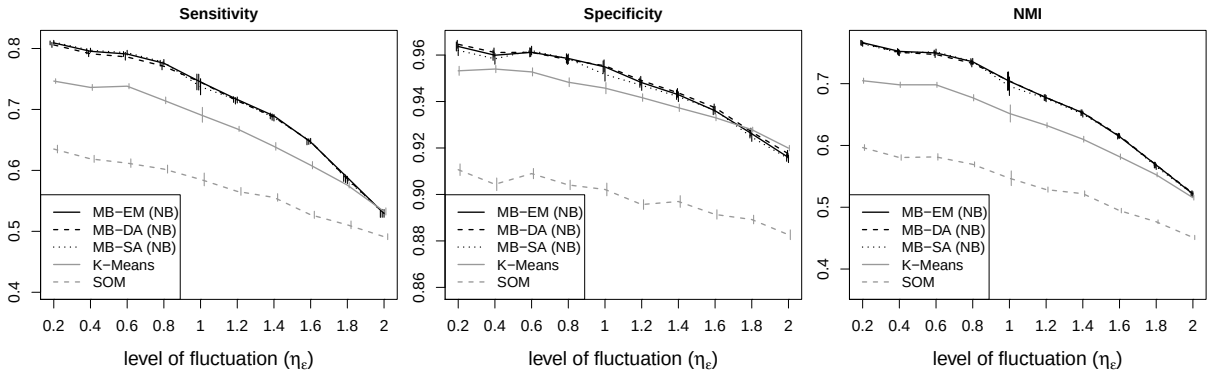
(b)

Figure 4.2: *Evaluate initialization of cluster centers.* The two methods for initialization for MB-EM algorithms were compared: using initialization with model-based algorithm (Algorithm 4.2) versus initialization with randomly picked objects(genes). For each parameter setting, results from 50 data sets were averaged and plotted on the line. The length of each vertical bar represents standard error.

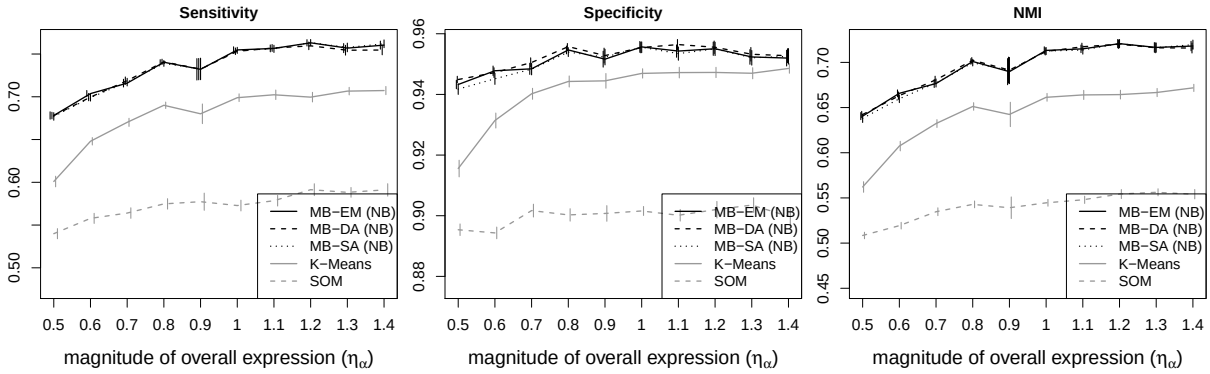
In Figure 4.2, we compared the initialization effects on the MB-EM clustering results. Our proposed model-based algorithm (Algorithm 4.2) and random initialization were examined. Though initialization using true cluster centers is not applicable in practice, we also included it in the comparison as a standard to evaluate the other two initialization methods. Figure

4.2 clearly illustrate that the model-based initialization performs much better than random initialization by giving higher sensitivity, specificity, and NMI for all parameter settings in simulation. In many cases, the model-based approach generated results similar to those when the true cluster centers were applied for initialization. Results for other simulation settings are presented in 4.A.1.

4.4.5 Comparison of Our Proposed Algorithms with Others



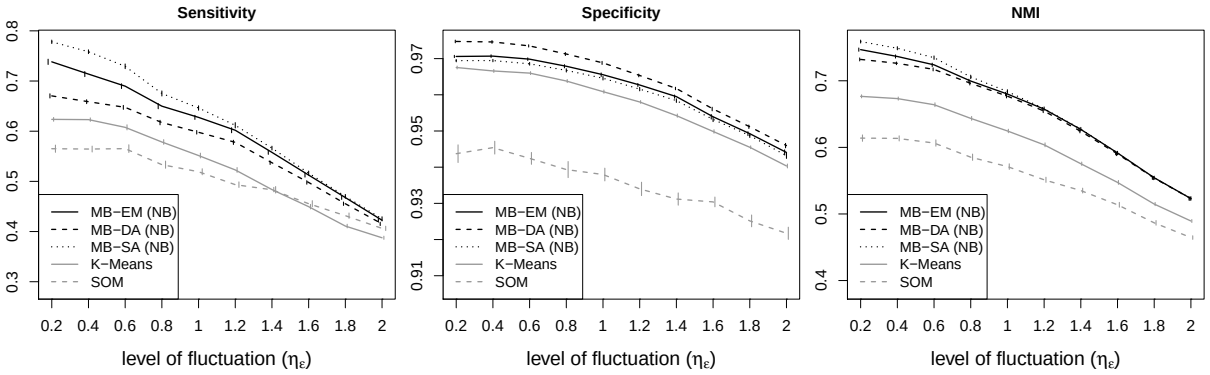
(a)



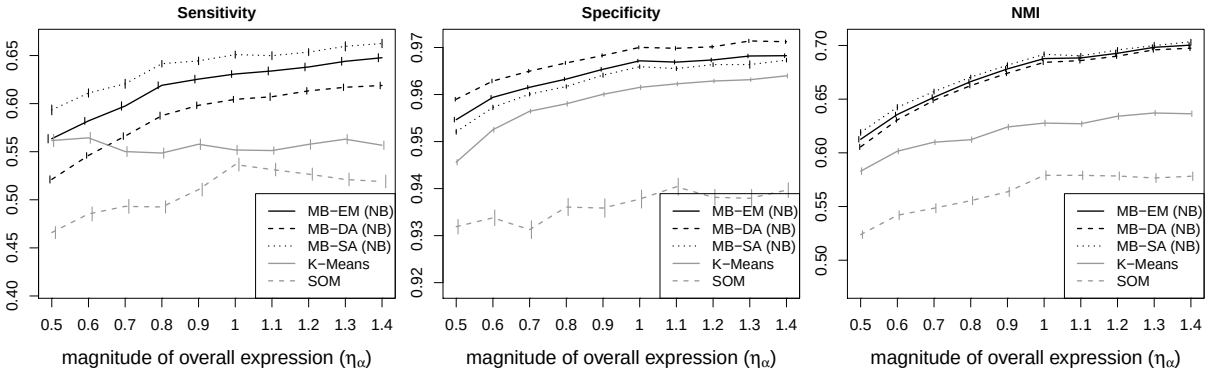
(b)

Figure 4.3: *Results of seven clusters from different clustering methods.* The model-based methods include EM, DA and SA algorithms, initialized by the same 7 cluster centers chosen by Algorithm 4.2. The non-MB methods include the standard K-means and SOM. For each parameter setting, results from 50 data sets were averaged and plotted on the line. The length of each vertical bar represents standard error.

We proposed EM algorithm (Algorithm 4.1) to perform model-based clustering. However, it is possible that the resulting partition from EM algorithm is not a global optimum. Hence, two



(a)



(b)

Figure 4.4: *Results of ten clusters from different clustering methods.* The model-based methods include EM, DA and SA algorithms initialized by the same ten cluster centers chosen by Algorithm 4.2. The non-MB methods include the standard K-means and SOM. For each parameter setting, results from 50 data sets were averaged and plotted on the line. The length of each vertical bar represents standard error.

stochastic versions, DA and SA algorithms, are described in section 4.3.3 to reduce such risk. In this section, we compare these slightly differing algorithms, while all three were initialized with the same set of cluster centers chosen by Algorithm 4.2. First, we did cluster analysis with the true number of clusters, $K = 7$. Figure 4.3 suggests that all three algorithms performs almost the same. We also analyzed the same data sets with $K = 10$. Interestingly, Figure 4.4 shows that the SA algorithm typically achieves the highest sensitivity while the DA algorithm gains in terms of specificity. If practitioners are more interested in specificity, getting groups of genes with similar profiles, then the SA algorithm is recommended. If separating genes with different profiles is more of concern, then DA algorithm can be applied.

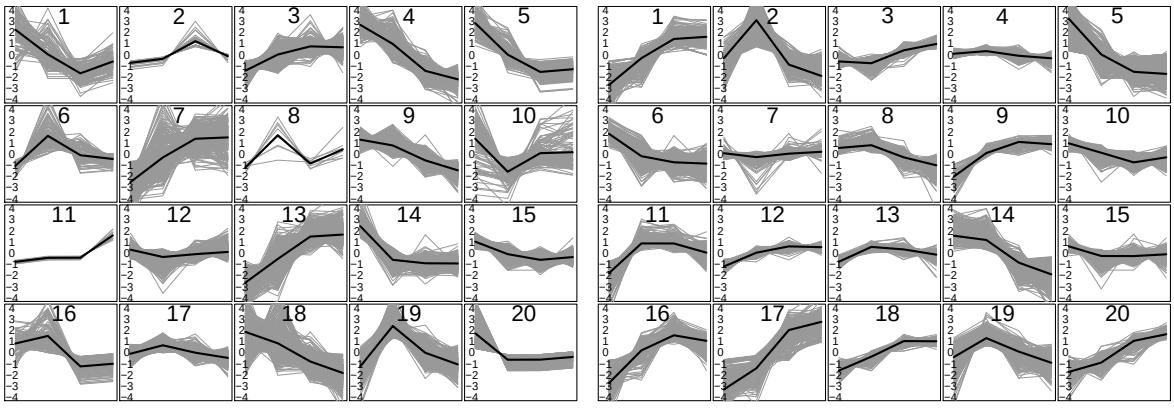
We also compared the proposed algorithms with K-means and self-organizing map (SOM), two methods that have been popularly applied to microarray analysis and can potentially be applied for RNA-seq data. To cluster gene expression profiles, K-means and SOM were applied to cluster the maximum likelihood estimates (MLE) of in the NB model. Plots in Figure 4.3, Figure 4.4 and 4.A.1 show that, evaluated by all three criteria, the model-based algorithms perform obviously better than K-means and even better than SOM. Note that our simulation settings include Poisson model, which is a special case when the dispersion parameter is set to be zero. We also did more simulations with Poisson model and the results are similar to what are shown here.

4.5 Real Data Analysis

Li et al. (2010) studied the maize leaf transcriptome using Illumina Genome Analyzer 2, one platform of NGS technologies. The dataset quantifies transcript abundance of four sections along a leaf developmental gradient, with two biological replicates for each section. Using generalized linear model analysis based on negative binomial distribution, we found that 12,631 genes were differentially expressed (DE) across the four sections. Li et al. (2010) normalized the count data by calculating the values of reads per kilobase of exon per million reads (RPKM), a popular quantification method proposed by Mortazavi et al. (2008). In this section, upon log-transform and, for each gene, mean-center the RPKM values, we applied both the K-means, which has been used in Li et al. (2010), and the SOM algorithms. We also present results from the model-based clustering algorithms based on NB model. The results show that our proposed method provides better clusters than both K-means and SOM algorithms.

First, we clustered the DE genes into $K = 20$ clusters with the same initial cluster centers chosen by Algorithm 4.2. Figure 4.5(a) and Figure 4.5(b) show the clusters given by K-means and MB-EM algorithm, respectively. Some clusters produced by the K-means method, e.g., cluster 7 and 18, contain genes with apparently different patterns of expression changes. In contrast, genes in the clusters given by the MB method show less variable expression patterns. The results from DA and SA algorithms look similar to Figure 4.5(b). This visual inspection of gene expression profiles for the clusters indicates that the model-based algorithms may work

better than K-means.

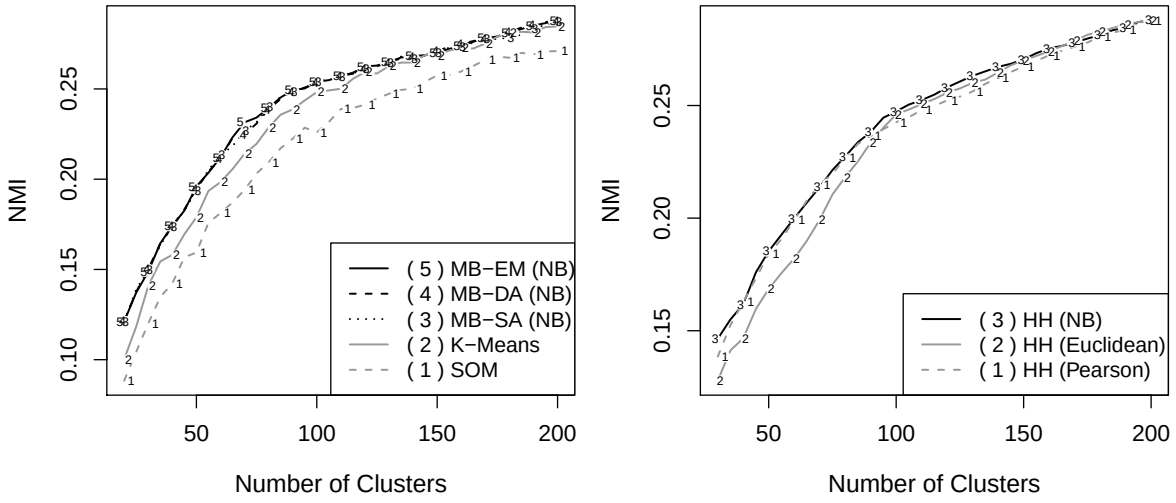


(a) K-means (b) MB-EM (NB)

Figure 4.5: (a) Result from K-means algorithm using Euclidean distance; (b) Result from EM algorithm based on NB model. The grey lines correspond to the estimated gene expression pattern estimated by method of moments, and the black lines plot the cluster centers.

In addition to the visual inspection, we also quantitatively compared different clustering algorithms by the NMI scores between clustering results and gene annoations. Gene annotations were obtained from Mapman as described in Li et al. (2010). Excluding categories that contain less than five or more than 500 genes, we ended with 126 non-overlapping categories with a total of 5075 genes. We expect that the genes within the same functional category have correlated expression patterns and thus more likely to be grouped together. So a clustering result can be evaluated by checking its concordance with the functional categories, where the concordance is measured by NMI. Furthermore, because these annotations are independent to the clustering processes, the evaluation is not biased toward any clustering method and data model.

We performed cluster analysis with $K = 10, 15, 20, \cdots, 200$ clusters. Figure 4.6(a) shows the NMI scores for all five methods, including SOM, K-means and the three model-based algorithms. The model-based algorithms outperform SOM and K-means for all K values. We also calculated the Bayesian information criterion (BIC) based on NB model. Not surprisingly, the results from model-based algorithms produced much smaller BIC than others (see 4.A.2). Another advantage of the model-based appraoches is that the Poisson or NB model can handle genes with low counts easily. When sequencing depth is low, there may be many genes with



(a) Model-based Clustering

(b) Hybrid-Hierarchical clustering

Figure 4.6: Clustering results for the maize data set. (a) We compared our proposed model-based algorithms (EM, DA and SA) with the K-means and SOM methods. (b) Model-based hybrid-hierarchical (MB-HH) is compared with hierarchical clustering based on Euclidean distance or Pearson correlation. They all start from the same set of 200 clusters

low counts or zero counts in some replicates or treatment groups. However, this will induce problems in the log-transformation which is typically done before applying K-means method.

We then applied the hybrid-hierarchical (HH) clustering as described in section 4.3.4, starting from $K_0 = 200$ clusters obtained from the MB-EM algorithm. We again employed hierarchical clustering using average linkage based on Euclidean distance or Pearson correlation starting from the same set of 200 clusters. Our proposed HH method generated higher NMI scores (Figure 4.6(b)) and lower BIC scores (see 4.A.2) than the other two hierarchical methods. The hierarchical structures for the MB-HH clusters are plotted in 4.A.2.

4.6 Conclusion

In this paper, we derived clustering algorithms based on Poisson and NB models that have been popularly used for RNA-seq data analysis. As explained in section 4.2, we recommend the Poisson model for data without biological replicates and NB model to handle data with biological replicates. We proposed an EM algorithm with model-based initialization, and show

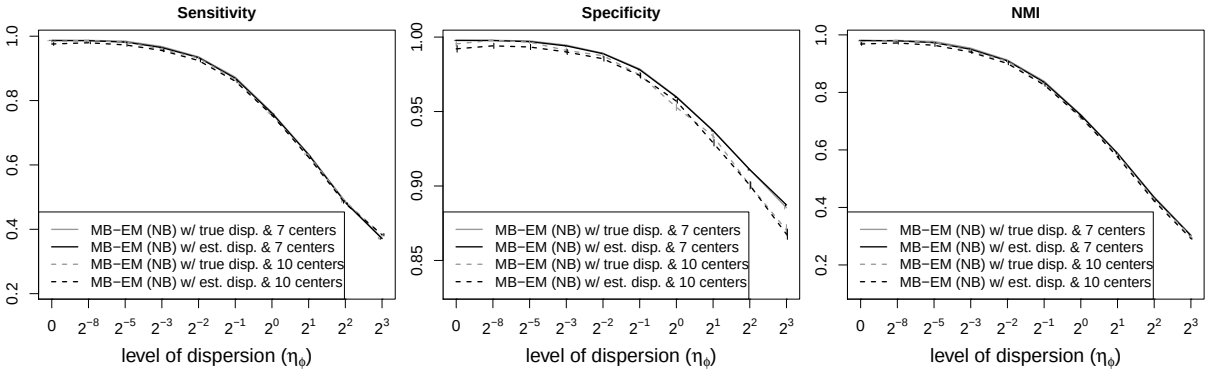
this initialization method greatly improves the performance of the EM clustering. We also introduced two stochastic versions of the EM algorithm and examined their performance. We demonstrated through both simulation studies and real data analysis that our proposed algorithms outperformed heuristic methods such as K-means and SOM, which have been popularly applied to cluster gene expressions from microarray and can also be applied to RNA-seq data.

We have developed an R package named `MBCluster.Seq` that implements our proposed algorithms. This R package provides fast computation and is publicly available at CRAN.

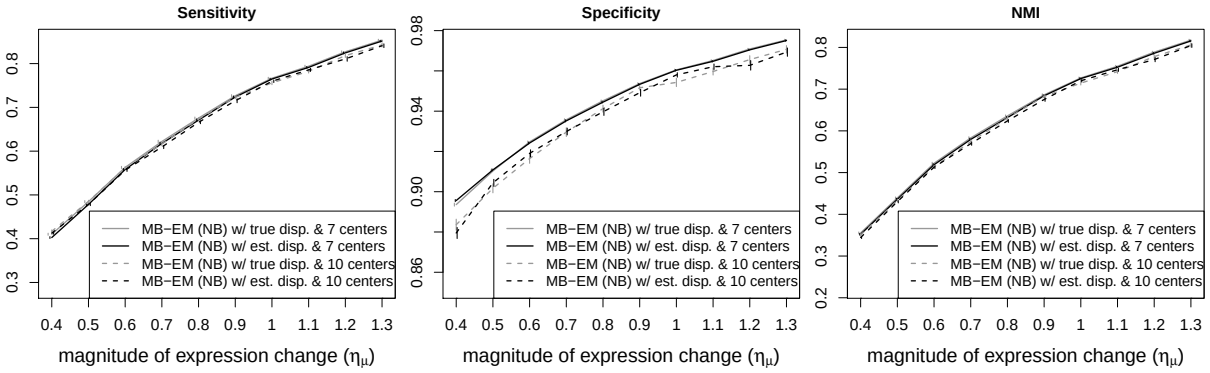
4.7 APPENDICES

4.A.1 Clustering Results for Simulation

The sensitivity, specificity and NMI scores are used to assess different clustering methods in the simulatios study in section 4.4

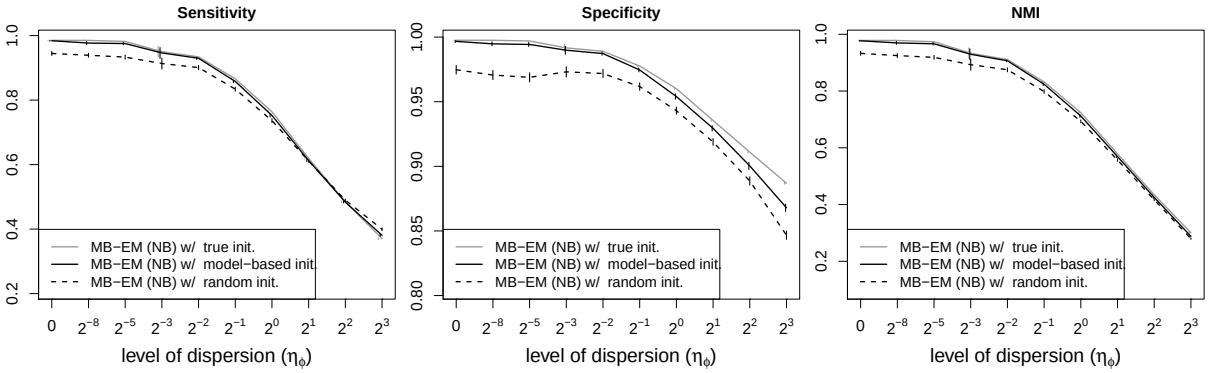


(a)

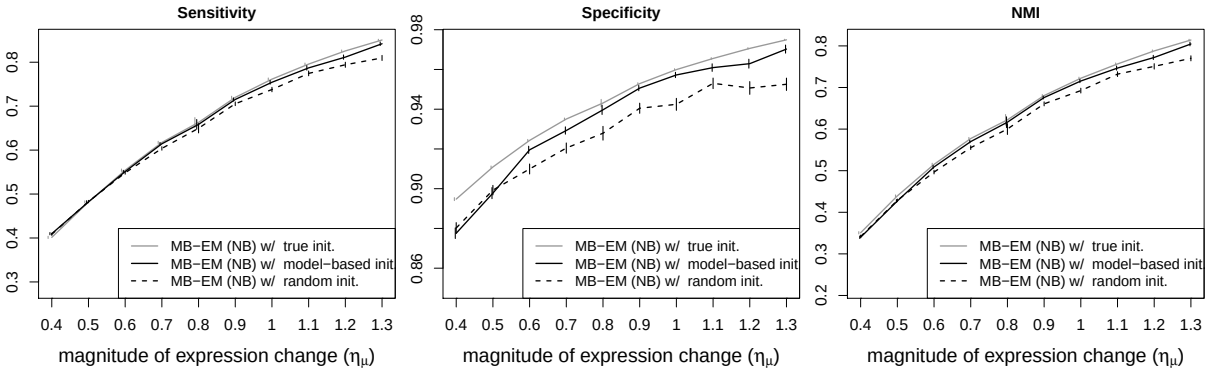


(b)

Figure 4.7: *Estimation of dispersion parameters.* Clustering results from the MB-EM algorithms using true and estimated dispersion parameters were compared. For each parameter setting, result from 50 data sets were averaged and plotted on the line. The length of vertical bars represents standard error. Solid lines are for results of 7 clusters and dashed lines are for results of 10 clusters.

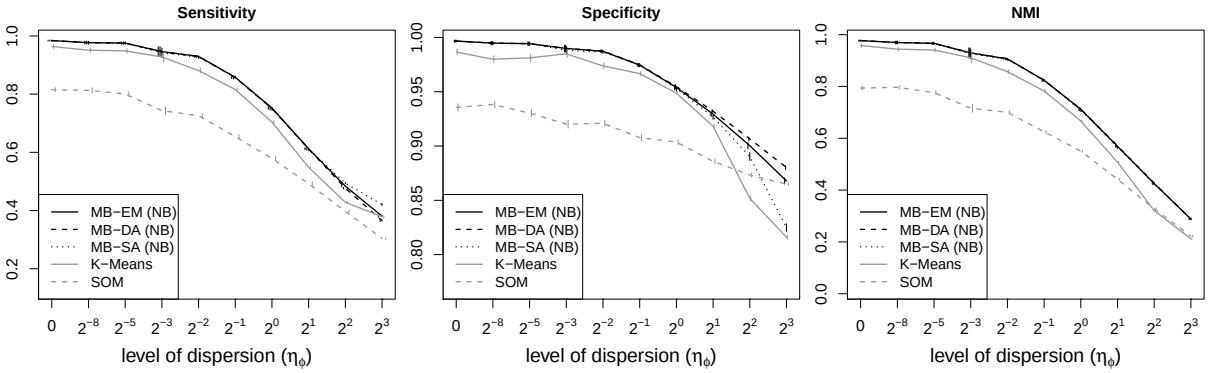


(a)

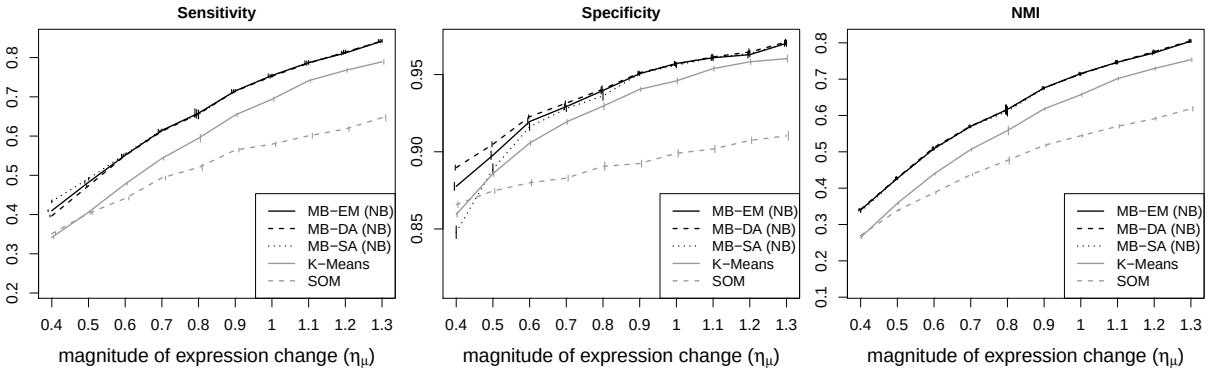


(b)

Figure 4.8: *Evaluate initialization of cluster centers.* The two methods for initialization for MB-EM algorithms were compared: using initialization with model-based ALGORITHM 2 versus initialization with randomly picked objects(genes). For each parameter setting, results from 50 data sets were averaged and plotted on the line. The length of each vertical bar represents standard error.

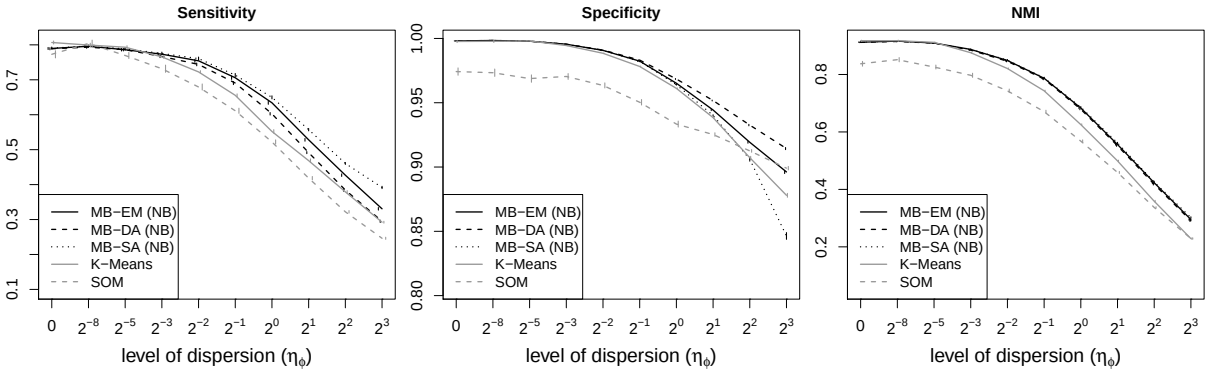


(a)

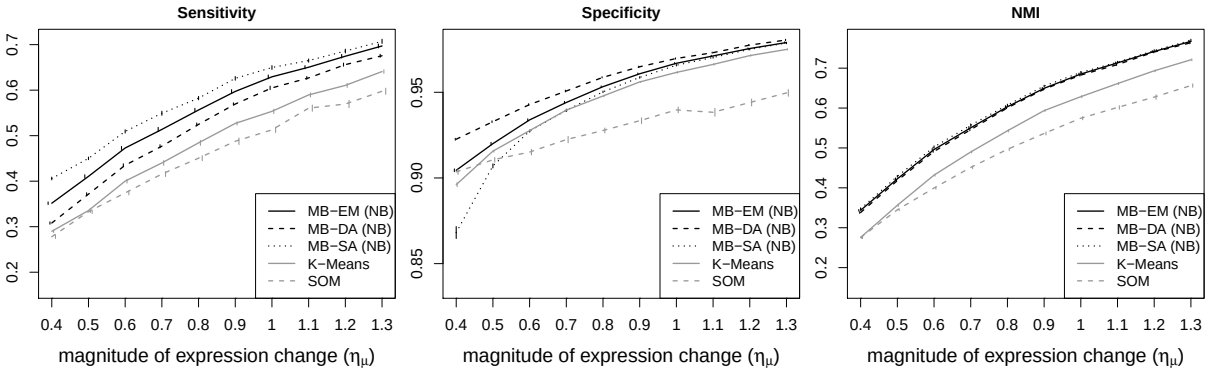


(b)

Figure 4.9: *Results of seven clusters from different clustering methods.* The model-based methods include EM, DA and SA algorithms initialized by the same 7 cluster centers chosen by Algorithm 2. The non-MB methods include the standard K-means and SOM. For each parameter setting, results from 50 data sets were averaged and plotted on the line. The length of each vertical bar represents standard error.



(a)

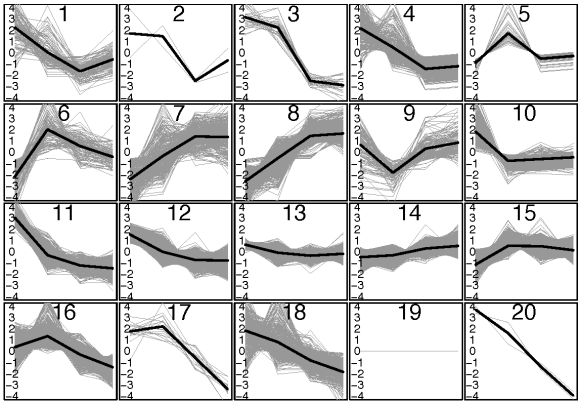


(b)

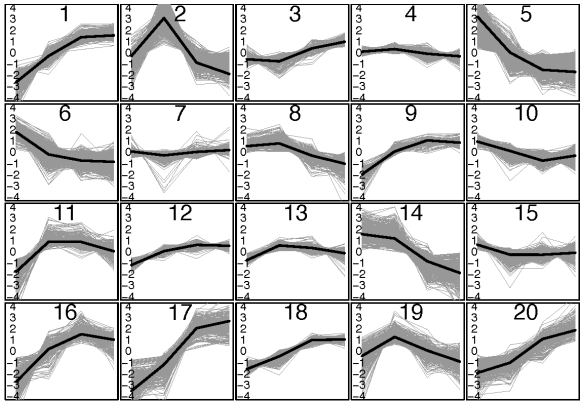
Figure 4.10: *Results of ten clusters from different clustering methods.* The model-based methods include EM, DA and SA algorithms initialized by the same ten cluster centers chosen by Algorithm 2. The non-MB methods include the standard K-means and SOM. For each parameter setting, results from 50 data sets were averaged and plotted on the line. The length of each vertical bar represents standard error.

4.A.2 Clustering Results for Real Data Analysis

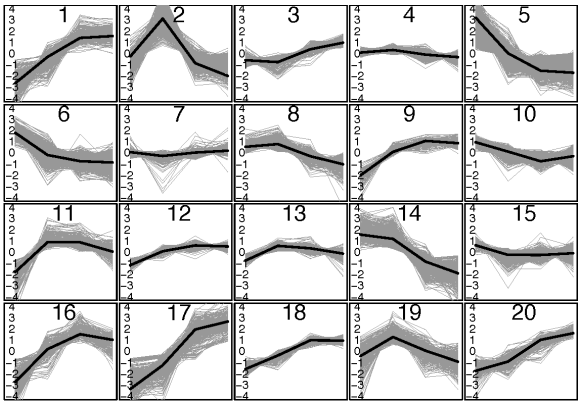
For the analysis of the maize data in section 4.5, the BIC scores for the clustering results are compared, and the tree structures of the hybrid-hierarchical clusters are plotted



(a) SOM

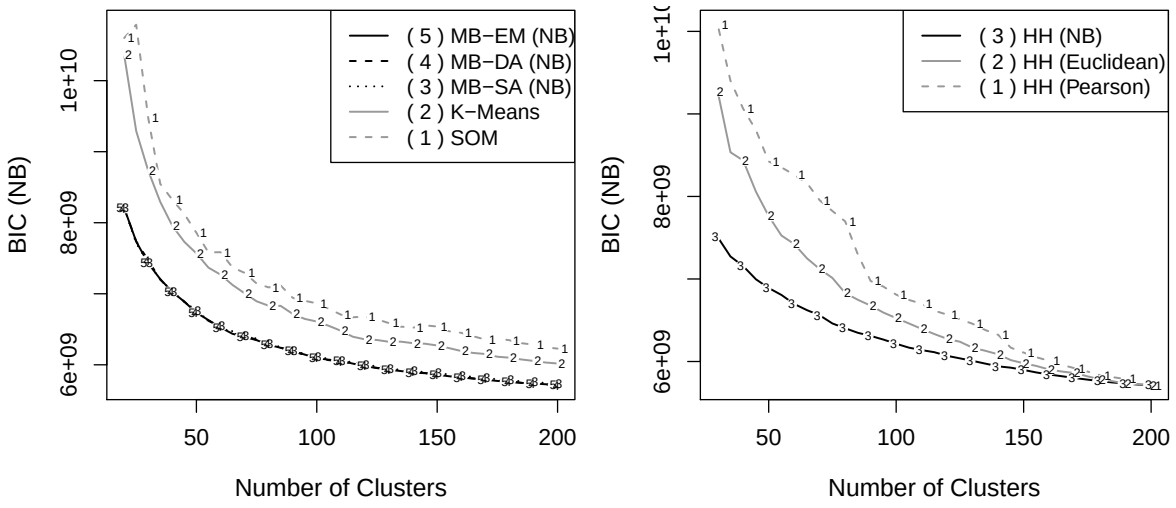


(b) MB-DA (NB)



(c) MB-SA (NB)

Figure 4.11: (a) Result from SOM algorithm using Euclidean distance; (b) Result from DA algorithm based on NB model; (c) Result from SA algorithm based on NB model. The grey lines correspond to the estimated gene expression pattern estimated by method of moments, and the black lines plot the cluster centers.



(a) Model-based Clustering

(b) Hybrid-Hierarchical clustering

Figure 4.12: Clustering results for the maize data set. (a) We compared our proposed model-based algorithms (EM, DA and SA) with the K-means and SOM methods. (b) Model-based hybrid-hierarchical (MB-HH) is compared with hierarchical clustering based on Euclidean distance or Pearson correlation. They all start from the same set of 200 clusters

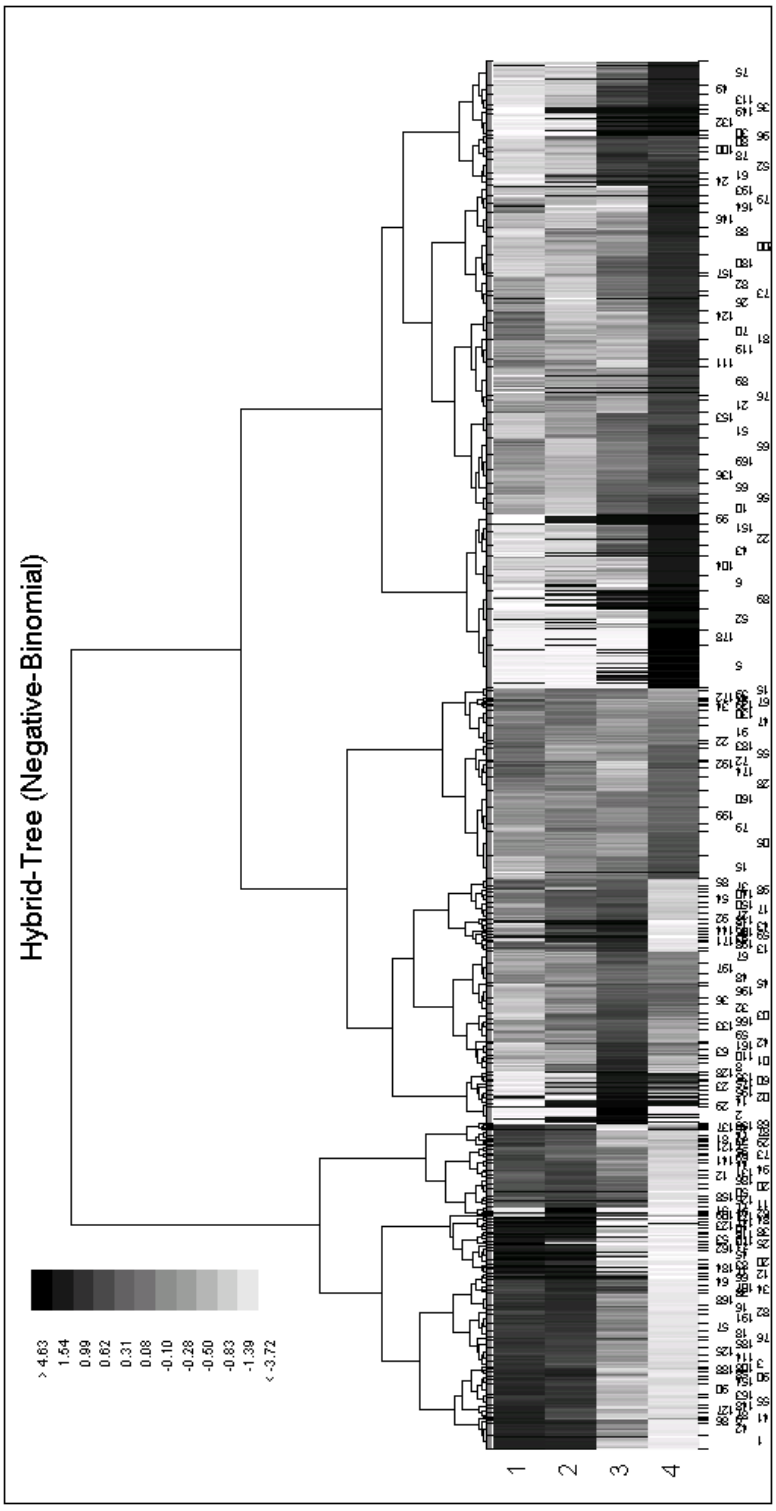


Figure 4.13: Tree structures of clusters from model-based hybrid-hierarchical clustering, using NB model

CHAPTER 5. General Discussion

We have provided a framework for building the approximated most average-powerful (AMAP) test that can be used to compare gene (Chapter 2) or exon (Chapter 3) expressions from RNA-seq data. Compared with previous work, the AMAP test enjoys great flexibility for null hypothesis settings and can deal with various questions depending on biological context. For example, it can be used for identifying differential expressions, detecting fold changes larger than a threshold and finding switch-like patterns using exon coverages, etc. We derived that, when the prior distribution is known, our test is the optimal in maximizing the average power while controlling false discovery rate (FDR). and this property was justified from intensive simulation studies. We also found a novel approach to accurately control FDR based on AMAP test statistics. For future research, we have noticed that the performance of AMAP test sometimes cannot match the theoretical MAP test, and might depend on how well we can estimate the dispersion parameters in the NB models and the prior distribution of gene/exon expressions. Hence more improvement might be achievable from better estimation. A hint is that Chapter 3 estimates the distribution of exon usages through a non-parametric approach, which is even more flexible and computationally efficient than the mixture distribution and the EM algorithm proposed for gene expression data in Chapter 2. So generalizing the non-parametric method of estimation to the later might be useful. Moreover, we have focused on comparing gene/exon expressions from two-treatment experiments, and it is desirable for us to search for efficient methodologies to analyze RNA-seq data from multi-treatment experiments.

The model-based clustering algorithm in Chapter 4 has been shown, through both simulation studies and real data analysis, to be able to outperform heuristic methods such as K-means and SOM. Like in many other clustering algorithms, a major question for our clustering strategy is that we need to predetermine the number of clusters, K, in the dataset, which is actually

unknown for RNA-seq data. Though the probability assumption employed by model-based clustering provides some convenience to deciding the K based on likelihood of the model, such as using the BIC criterion, sometimes, we did not observed a 'turning point' of BIC or other criterions in our real data analysis. Hence in practice, we still need to choose the K partially based on some prior knowledge, experiences, or other techniques. So a more plausible method of choosing the cluster number is still desirable, and we will try to settle this question in the future.

BIBLIOGRAPHY

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Proceedings* **11,** R106.

Anders, S., Reyes, A. and Huber, W. 2012, *Detecting differential usage of exons from RNA-Seq data*. Nature Precedings, hdl:10101/npre.2012.6837.1

Anderson, T.W. (1962). On the Distribution of the Two-Sample Cramervon Mises Criterion *The Annals of Mathematical Statistics* **33 (3)**: 11481159.

Arthur, D. and Vassilvitskii, S. (2007). K-means++: the advantages of careful seeding *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms,* 1027–1035.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B* **57,** 289-300.

Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing *Annual Review of Biochemistry* **72**:291-336

Bloom, J., Khan, Z., Kruglyak, L., Singh, M. and Caudy, A. (2009). Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarray. *BMC Genomics* **10,** 221.

Booth, J.G., Casella, G. and Hobert, J.P. (2008). Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society: Series B* **70** 119–139.

Bullard, J.H., Purdom, E.A., Hansen, K.D. and Dudoit, S. (2010). Evaluation of Statistical

Methods for Normalization and Differential Expression in mRNA-Seq Experiments. *BMC Bioinformatics* **11,** 94.

Caceres J.F., Kornblihtt A.R., 2002, Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet* **18(4)**:186-193.

Calarco J.A., Superina S., O'Hanlon D., et al. 2009, Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. *Cell* **138(5)**:898-910.

Casella, G. and Berger, R.L. (2002) Statistical Inference *Thomson Learning*

Celeux, G. and Govaert, G. (1992). EA Classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis,* **14** 315–332.

Chen, L., Hung, H. and Chen, C. (2007). Maximum Average-Power(MAP) Tests. *Commun Stat. A-Theor.* **36,** 2237-2249.

Cooper T.A. (2005). Alternative splicing regulation impacts heart development. *Cell*, **120(1)**:1-2.

Covshoff, S., Majeran, W., Liu P., Kolkman, J. M., van Wijk, K. J. and Brutnell, T.B. (2008). De-regulation of maize C4 photosynthetic development in a mesophyll cell defective. *Plant Physiology* **146,** 1469-1481.

Cox, D. R. and Reid, N. (1987). Parameter Orthogonality and Approximate Conditional Inference *Journal of the Royal Statistical Society. Series B (Methodological)* **49(1)**:1-39.

Fraley, C. (1999). Algorithms for model-based gaussian hierarchical clustering. *SIAM Journal on Scientific Computing* **20(1)** 270–281.

Fraley, C. and Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97** 611–631.

Hall, L.O., Özyurt, I.B. and Bezdek, J.C. (1999). Clustering with a genetically optimized approach. *IEEE Transactions On Evolutionary Computation* **3** 103–112.

Hardcastle, T.J. and Kelly, K.A. (2010). baySeq: Empirical Bayesian Methods for Identifying Differential Gene Expression in Sequence Count Data. *BMC Bioinformatics* **11,** 422.

Hwang, J.T.G. and Liu, P. (2010). Optimal Tests Shrinking Both Means and Variances Applicable to Microarray Data Analysis. *Statistical Applications in Genetics and Molecular Biology* **9,** Issue 1, Article 36.

Ji, H., Jiang, H., Ma, W., Johnson, D. S., Myers, R. M. and Wong, W. H. (2008). Statistical Inferences for Isoform Expression in RNA-Seq. *Bioinformatics* **25,** 1026-1032.

Johnson, J. M. et al. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**:21412144

Keren H, Lev-Maor G, Ast G, 2010, Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet*, **11(5)**:345-355.

Kvam, V., Liu, P. and Si, Y. (2012). A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American Journal of Botany* **99(2),** 248-256.

Lander, E. S. et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**:860921

Li, J., Jiang, H., Wong, W.H. (2010) Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology* **11**:R50

Li, J., and Tibshirani, R.. (2009). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research* **Published online**

Li, P., Ponnala, L., Gandotra, N., Wang, L., Si, Y., Tausta, S.L. et al. (2010). The developmental dynamics of the maize leaf transcriptome. *Nature Genetics* **42,** 1060-1067.

Li, J., Witten, D. M., Johnstone, I. M., and Tibshirani, R.. (2009). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatics* **Published online**

MacCarthy, D.J. and Smyth, G.K. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* **25,** 765-771.

Marguerat, S., Wilhelm, B.T. and Bähler, J. (2008). Next-Generation Sequencing: Applications beyond Genomes *Biochemical Society Transactions* **36** 1091–1096.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18,** 1509-1517.

Meila, M. and Heckerman, D. (2001). An experimental comparison of model-based clustering methods. *Machine Learning* **42** 9–29.

Metzker, M.L. (2010). Sequencing technologies – the next generation *Nature Reviews Genetics* **11** 31–46.

Mortazavi, A., Williams, B. A., McCue, K, Schaeffer, L. and Wold, B. (2008). Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq. *Nature Methods* **5,** 621-628.

Nelder, J.A. (2000). Quasi-likelihood and pseudo-likelihood are not the same thing. *Journal of Applied Statistics* **27,** 1007-1011.

Robinson, M. D. and Young, M. D. (2010). From RNA-seq reads to differential expression results *Genome Biology* **11,** 220.

Park, H.S., Yoo, S.H. and Cho, S.B. Evolutionary fuzzy clustering algorithm with knowledge-based evaluation and applications for gene expression profiling (2005). *Journal of Computational and Theoretical Nanoscience* **2** 1–10.

Peart, M.J., Smyth, G.K., van Laar, R.K., Bowtell, D.D., Richon, V.M., Marks, P.A. et al. (2005). Identification and functional significance of genes regulated by structurally different histone deacetylase inhibitors. *Proceedings of the National Academy of Sciences, USA* **102,** 3697-3702.

Pickrell, J., Marioni, J., Pai, A., Degner, J., Engelhardt, B., Nkadori, E. et al. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768-772.

Priness, I., Maimon, O. and Ben-Gal, I. (2007). Evaluation of gene-expression clustering via mutual information distance measure *BMC Bioinformatics* **8** 111.

Ressom, H., Wang, D. and Natarajan, P. (2003). Clustering gene expression data using adaptive double self-organizing map. *Physiological Genomics* **14** 35–46.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T. et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing *Nature Methods* **4,** 651-657.

Robinson, M.D. and Oshlack, A. (2010). A Scaling Normalization Method for Differential Expression Analysis of RNA-seq Data. *Genome Biology* **11,** R25.

Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance *Bioinformatics* **23,** 2881.

Robinson, M.D. and Smyth, G.K. (2008). Small-Sample Estimation of Negative Binomial Dispersion, with Applications to SAGE Data. *Biostatistics* **9,** 321-332.

Rose, K. (1998). Deterministic annealing for clustering, compression, classification, regression, and related optimization Problems. *Proceeding of the IEEE* **86** 2210–2239.

Shen, S., Park, J.W., Huang J., Dittmar K.A., Lu Z, Zhou, Q., Carstens, R.P. and Xing, Y. (2012) MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data *Nucleic Acids Research* **1-13** doi:10.1093/nar/gkr1291.

Shendure J, Ji H (2008) Next-generation RNA sequencing. *Nature Biotechnology* **26**: 2514-2521

Simonoff, J.S. (1996) Smoothing Methods in Statistics *Springer*

Smyth, G. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray. *Statistical Applications in Genetics and Molecular Biology* **3** A.3

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T.A., and Soreq, H. (2005) Function of alternative splicing. *Gene* **344**: 120.

Storey, J.D. and Tibshirani, R. (2003). Statistical Significance for Genome-Wide Studies. *Proceedings of the National Academy of Sciences* **100,** 9440-9445.

Storey, J.D. (2007). The Optimal Discovery Procedure: a New Approach to Simultaneous Significance Testing. *Journal of the Royal Statistical Society: Series B* **69,** 347-368.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining partitions. *Journal of Machine Learning Research* **3** 583–617.

Sultan, M., Schulz, M. and Richard, H. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321,** 956-960.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **96(6)** 2907–2912.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response *Proceedings of the National Academy of Sciences of the United States of America* **98,** 5115-21.

Vaithyanathan, S. and Dom, B. (2000). Model-based hierarchical clustering. *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence* 599–608.

Vardhanabhuti, S., Li, M. and Li, H. 2011, *A Hierarchical Bayesian Model for Estimating and Inferring Differential Isoform Expression for Multi-sample RNA-Seq Data.* Stat Biosci, DOI 10.1007/s12561-011-9052-3.

Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10** 53–67.

Wang, L., Li, P. and Brutnell, T. P. (2010). Exploring plant transcriptomes using ultra high-throughput sequencing. *Briefings in Functional Genomics* **9** 118–128.

Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**:470476.

Wang L, Si Y, Dedow LK, Shao Y, Liu P, Brutnell TP, 2010, A low-cost library construction protocol and data analysis pipeline for Illumina-based strand-specific multiplex RNA-seq. *PLoS One* **6(10)**:e26426.

Woodard, D.B. and Goldszmidt, M. (2011). Model-based clustering for online crisis identification in distributed computing. *Journal of the American Statistical Association* **106(493)** 49–60.

Wu, Z., Jenkins, B.D., Rynearson, T.A., Dyhrman, S.T., Saito, M.A. and Mercier, M. (2010). Empirical bayes analysis of sequencing-based transcriptal profilling without replicates. *BMC Bioinformatics* **11:** 564.

Xiao, X., Dow, E.R., Eberhart, R., Miled, Z.B., and Oppelt, R.J. (2003) Gene Clustering Using Self-Organizing Maps and Particle Swarm Optimization *International Parallel and Distributed Processing Symposium (IPDPS'03)* p154b

Xing,Y. and Lee,C.J. (2005) Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS Genetics* **1**: e34

Yeung, K.Y., Fraley, C., Murua, A., Faftery, A.E., and Ruzzo, W.L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17(10),** 977–987.

Zhang XC and Gassmann W, 2007, Alternative splicing and mRNA levels of the disease resistance gene RPS4 are induced during defense responses. *Plant Physiol* **145(4)**:1577-1587.

Zhong, S. and Ghosh, J. (2003). A unified framework for model-based clustering *Journal of Machine Learning Research* **4** 1001–1037.